John von Neumann Institute for Computing

Ab Initio Methods for Electron Correlation in Molecules

Peter Knowles, Martin Schütz, and Hans-Joachim Werner

published in

Modern Methods and Algorithms of Quantum Chemistry,
Proceedings, Second Edition, J. Grotendorst (Ed.),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **3**, ISBN 3-00-005834-6, pp. 97-179, 2000.

http://www.fz-juelich.de/nic-series/

AB INITIO METHODS FOR ELECTRON CORRELATION IN MOLECULES

PETER KNOWLES
School of Chemistry
University of Birmingham
Edgbaston
Birmingham, B15 2TT
United Kingdom
E-mail: P.J.Knowles@bham.ac.uk

MARTIN SCHÜTZ AND HANS-JOACHIM WERNER
Institute for Theoretical Chemistry
University of Stuttgart
Pfaffenwaldring 55
70569 Stuttgart
Germany
E-mail: {schuetz, werner}@theochem.uni-stuttgart.de

Reliable *ab initio* electronic structure calculations require high-level treatment of electron correlation effects. For molecules in electronic ground states, single-reference correlation methods, which are based on the Hartree-Fock self-consistent field (SCF) wavefunctions as zeroth order approximation, are usually sufficient. Møller-Plesset perturbation theory up to fourth order (MP2-MP4) and coupled-cluster methods with all single and double excitations followed by a perturbative treatment of triple excitations [CCSD(T)] are the most popular single-reference methods. All of these approaches can also be formulated in a local framework which gives a demand on computational resources that scales only linearly with system size; they can also be carried out using integral-direct techniques, that avoid the storage of large numbers of two electron integrals by recomputing them on demand. For computing electronically excited states or global potential energy functions, multiconfiguration self-consistent field (MCSCF) wavefunctions are required for a qualitatively correct representation of the wavefunction. The major part of dynamical electron correlation effects can then be accounted for by subsequent multireference correlation treatments, in which a large number of single and double excitations relative to the MCSCF reference configurations are taken into account. In multireference configuration interaction (MRCI) calculations the expansion coefficients are determined variationally. Alternatively, the coefficients can be obtained by first-order perturbation theory, and the energy be evaluated to second (MRPT2) or third (MRPT3) order. These lecture notes give a short review of all these methods.

1 Introduction

1.1 Electron correlation and the configuration interaction method

Hartree-Fock Self-Consistent Field (SCF) Theory enjoys considerable success in the first-principles determination of molecular electronic wavefunctions and properties. However, there are important situations where the underlying assumption of molecular orbital theory, that the electronic wavefunction can be approximated by an antisymmetrized product of orbitals, breaks down. There are still further situations where SCF does provide a reasonable qualitative description, but fails to

predict energetics to desired accuracy. We explore here the deficiencies of Hartree-Fock, and survey the various techniques available for going beyond SCF.

Hartree-Fock is a mean field theory, in which each electron has its own wavefunction (orbital), which in turn obeys an effective 1-electron Schrödinger equation. The effective hamiltonian (Fock operator) contains the average field (Coulomb and exchange) of all other electrons in the system. The total electronic wavefunction for the molecule, ignoring complications introduced by the Pauli principle, is a simple product of the orbitals. Following the Born interpretation of wavefunctions, this implies that if $P(r_1, r_2)$ is the probability density for finding electrons labelled 1 and 2 in regions of space around r_1 and r_2 respectively,

$$P(r_1, r_2) = P(r_1)P(r_2) \tag{1}$$

i.e., the probability density for a given electron is independent of the positions of all others.

In reality, however, the motions of electrons are more intimately correlated. Because of the direct Coulomb repulsion of electrons, the instantaneous position of electron 2 forms the centre of a region in space which electron 1 will avoid. This avoidance is more than that caused by the mean field, and is local; if electron 2 changes position, the Coulomb hole for electron 1 moves with it. In contrast, in the mean-field theory, electron 1 has no knowledge of the instantaneous position of 2, only its average value, and thus motions are uncorrelated, and there is no depletion in $P(r_1, r_2)$ near $r_1 = r_2$.

The effects of neglecting electron correlation in Hartree-Fock are spectacularly illustrated when one attempts to compute complete potential curves for diatomic molecules using SCF. Figure 1 shows potential curves for H_2 from both a very accurate calculation and from Hartree-Fock. It is seen that the spin-restricted Hartree-Fock (RHF) approximation breaks down as dissociation is reached, predicting energies which are much too high, and a potential curve characteristic of the interaction of ions rather than neutral atoms. The RHF wavefunction for the $X^1\Sigma_g^+$ ground state of H_2 takes the form

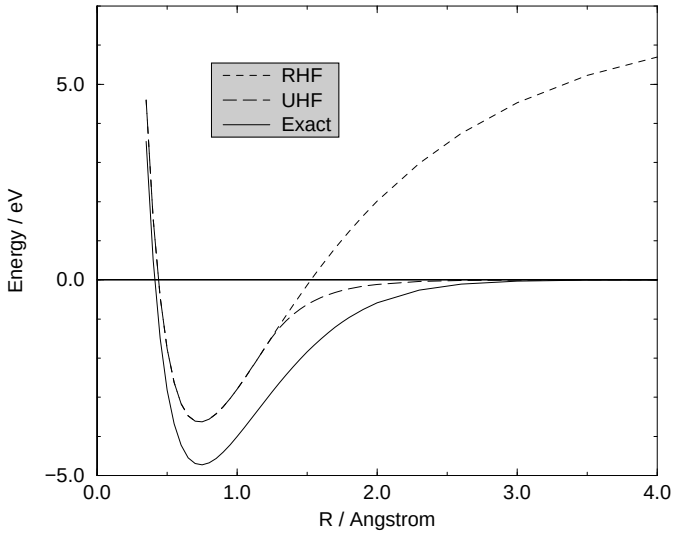
$$\Psi_{\mathrm{X}} = \hat{\mathcal{A}}\sigma_g^{\alpha}(1)\sigma_g^{\beta}(2) \tag{2}$$

where $\hat{\mathcal{A}}$ is the antisymmetrizing operator, α and β are the usual one-electron spin functions, and the bonding orbital $\sigma_g = Z_{\sigma_g}(\chi_{\mathrm{A}} + \chi_{\mathrm{B}})$, with χ_{A} an s-like orbital centred on atom A, and Z_{σ_g} a normalization constant. As the atoms become infinitely separated, $\chi_{\mathrm{A}} \sim 1s_{\mathrm{A}}$, $Z_{\sigma_g} \sim \frac{1}{\sqrt{2}}$ and thus

$$\Psi_{\mathrm{X}} \sim \frac{1}{2}\hat{\mathcal{A}}\left(1s_{\mathrm{A}}^{\alpha}1s_{\mathrm{B}}^{\beta} + 1s_{\mathrm{B}}^{\alpha}1s_{\mathrm{A}}^{\beta} + 1s_{\mathrm{A}}^{\alpha}1s_{\mathrm{A}}^{\beta} + 1s_{\mathrm{B}}^{\alpha}1s_{\mathrm{B}}^{\beta}\right) \tag{3}$$

The first two terms are direct products of neutral ^2S hydrogen atom wavefunctions on the two atoms A and B, as desired. However, the last two terms describe a spurious $H^+ \ldots H^-$ pair. The overall energy of this unphysical wavefunction exceeds the energy of two hydrogen atoms by half the difference of the ionization energy and electron affinity of H (i.e., 6.4 eV), and at long range the potential energy curve has an unphysical ionic R^{-1} behaviour.

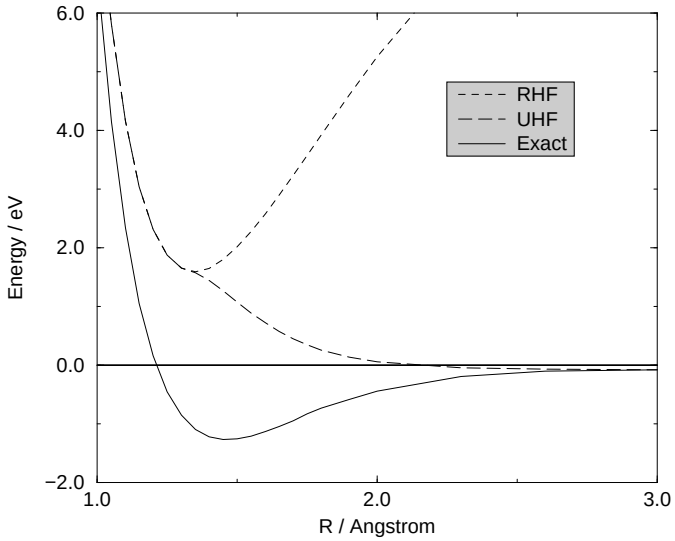
Figure 1. Potential Energy Curves for H_2



The failure of RHF for this example can be easily understood in terms of electron correlation. At long internuclear separations, if one electron is located near atom A, the other will on physical grounds be found close to atom B. This correlation is reflected in the exact wavefunction, which is asymptotically the product of hydrogenic orbitals on the two nuclei. In contrast, within the Hartree-Fock framework, each electron is made to experience only the average effect of the other. Since in RHF, the two electrons are constrained to be in the same spatial orbital, this σ_g orbital will be symmetrical between the atoms, and thus each electron has equal probability of being on A or B, irrespective of the position of the other electron. The possibility of both electrons being on the same atom is not excluded, as reflected in the ionic terms in the RHF wavefunction (3).

In the case of H_2, and in fact for a number of other dissociating molecules, Hartree-Fock theory can give correct behaviour provided the restriction to identical spatial orbitals for α and β spin is relaxed. The Unrestricted HF (UHF) wavefunction for H_2 is identical to RHF at short bond lengths, but when the two atoms are separated, it becomes variationally advantageous for the α and β spin orbitals to localize on different hydrogen atoms. In this way, a correct asymptotic energy is obtained, as seen in Figure 1. However, the wavefunction can never be identical to the exact wavefunction. Asymptotically, the UHF wavefunction is either $\hat{\mathcal{A}}1s_{\mathrm{A}}^{\alpha}1s_{\mathrm{B}}^{\beta}$ or $\hat{\mathcal{A}}1s_{\mathrm{B}}^{\alpha}1s_{\mathrm{A}}^{\beta}$, whereas the true wavefunction is the sum of these two degenerate determinants. Although the energy is unaffected, the UHF wavefunction is not an eigenfunction of the spin-squared operator $\hat{S}^2$, being an unphysical mixture of singlet and triplet states. This spin contamination is displeasing, and

Figure 2. Potential Energy Curves for F_2



can have serious undesirable effects. In the case of the H_2 UHF potential curve, at the point where UHF and RHF diverge, the curve is discontinuous in its second derivative. For more advanced correlation methods which build on UHF, spin contamination has a disastrous effect[1,2]. In the case of F_2 (Figure 2), UHF does not repair the inability of RHF to give an energy at equilibrium geometry which is lower than at dissociation, and as a consequence the UHF potential curve is purely repulsive. For all these reasons, the use of UHF is becoming increasingly rare.

1.2 Long-range correlation — Molecular Dissociation

In order to understand a theory which goes beyond the inability of RHF to describe dissociation, we examine first of all an excited $^1\Sigma_g^+$ state of H_2 for which the RHF wavefunction takes the form

$$\Psi_{\mathrm{E}} = \hat{\mathcal{A}}\sigma_u^{\alpha}(1)\sigma_u^{\beta}(2) \tag{4}$$

and where we now have two electrons in the antibonding orbital $\sigma_u = Z_{\sigma_u}(\chi_{\mathrm{A}} - \chi_{\mathrm{B}})$. Asymptotically, this becomes

$$\Psi_{\mathrm{E}} \sim \tfrac{1}{2}\hat{\mathcal{A}}\left(1s_{\mathrm{A}}^{\alpha}1s_{\mathrm{B}}^{\beta} + 1s_{\mathrm{B}}^{\alpha}1s_{\mathrm{A}}^{\beta} - 1s_{\mathrm{A}}^{\alpha}1s_{\mathrm{A}}^{\beta} - 1s_{\mathrm{B}}^{\alpha}1s_{\mathrm{B}}^{\beta}\right) \tag{5}$$

This wavefunction also contains an unphysical mixture of covalent and ionic terms. However, we observe that it is possible to construct purely ionic or purely covalent wavefunctions by taking a linear combination of Ψ_{X} and Ψ_{E}. In $\Psi_{\mathrm{X}} - \Psi_{\mathrm{E}} = \sigma_g^2 - \sigma_u^2$, the ionic terms cancel exactly, and the correct asymptotic wavefunction is obtained.

This is an example of *configuration interaction* (CI), whereby the wavefunction is considered as being a mixture of several Slater determinants. For H_2 at general internuclear separations, the form of the CI wavefunction is

$$\Psi = c_{\mathrm{X}}\Psi_{\mathrm{X}} + c_{\mathrm{E}}\Psi_{\mathrm{E}} \tag{6}$$

and the coefficients specifying this linear combination must be allowed to vary, since it is known that near equilibrium, the RHF wavefunction is already a good approximation. Thus the best wavefunction near equilibrium will have $c_{\mathrm{X}} \simeq 1$ and c_{E} small, in contrast to their asymptotic values of $\frac{1}{\sqrt{2}}$ and $-\frac{1}{\sqrt{2}}$.

In general, in the standard CI method, the *variational principle* is used to determine the CI coefficients. For any approximate wavefunction, the Rayleigh quotient

$$\overline{E} = \frac{\langle\Psi|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} \tag{7}$$

is an upper bound to the exact ground-state energy E, i.e., $\overline{E} \geq E$. Variational methods proceed by assuming that the best wavefunction will be the one which gives the lowest, i.e. minimum, $\overline{E}$. In the specific case of a linear expansion, as in CI, i.e.,

$$\Psi = \sum_I c_I \Phi_I \tag{8}$$

minimising $\overline{E}$ is equivalent to finding the lowest eigensolution of the *hamiltonian matrix* $\mathbf{H}$, whose elements are the integrals

$$H_{IJ} = \langle\Phi_I|\hat{H}|\Phi_J\rangle \ , \tag{9}$$

i.e. one needs to solve

$$\mathbf{Hc} = \overline{E}\mathbf{c} \tag{10}$$

with the minimum Rayleigh quotient $\overline{E}$ appearing as the eigenvalue. The linear ansatz allows also the calculation of approximations to excited states, through the Hylleraas-Undheim-MacDonald theorem, which states that the n-th eigenvalue is an upper bound to the exact energy of the $(n-1)$-th excited state. Finding the lowest few eigensolutions of a symmetric matrix is a well-studied problem; for the diagonally-dominant hamiltonian matrices invariably arising in molecular CI, algorithms exist[3] which will converge in around ten iterations, each of which requires the evaluation of the action of the hamiltonian matrix on some trial vector, i.e.,

$$v_I = \sum_J H_{IJ}c_J \ . \tag{11}$$

This feature allows the solution of CI problems of very large dimensions; because $\mathbf{H}$ is often extremely sparse, forming $\mathbf{H}\cdot\mathbf{c}$ is much easier than forming the matrix itself, and the limiting factor is the availability of memory to store $\mathbf{c}$ and $\mathbf{v}$. Calculations with more than 10^9 configurations have been carried out in this way.

1.3 Short-range correlation — the Interelectronic Cusp

Although consideration of electron correlation is clearly vital for the proper description of molecules closed to dissociation, it also has important implications in situations where Hartree-Fock is a reasonable approximation. Since the hamiltonian operator contains r_{ij}^{-1}, the inverse distance between two electrons, the nature of the electronic wavefunction in regions close to $r_{ij} = 0$ will have a strong effect on the energy.

We will consider initially the helium atom, for which the hamiltonian is

$$\hat{H} = -\tfrac{1}{2}\nabla_1^2 - \tfrac{1}{2}\nabla_2^2 - \frac{2}{r_1} - \frac{2}{r_2} + \frac{1}{r_{12}}\,. \tag{12}$$

The electronic wavefunction will satisfy Schrödinger's equation

$$\hat{H}\Psi(\mathbf{r}_1, \mathbf{r}_2) = E\Psi(\mathbf{r}_1, \mathbf{r}_2) \tag{13}$$

at all points in six-dimensional space. We note that close to $r_{12} = 0$ there is a paradox; the left hand side of (13) apparently becomes infinite, because of the $1/r_{12}$ Coulomb singularity, whereas E is constant, and so the right hand side is well behaved. The local energy $\hat{H}\Psi/\Psi$ cannot have singularities since it is constant, and the inescapable conclusion is that there must be an additional singularity in the left hand side of (13) which exactly cancels $1/r_{12}$ close to $r_{12} = 0$. Since the electrons are not necessarily close to a nucleus, the only candidate for this cancelling term is the kinetic energy. It is convenient to transform to centre-of-mass and relative coordinates,

$$\mathbf{R} = \tfrac{1}{2}\left(\mathbf{r}_2 + \mathbf{r}_1\right)\,; \qquad \mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1\,, \tag{14}$$

in which the hamiltonian becomes

$$\hat{H} = -\tfrac{1}{4}\nabla_{\mathbf{R}}^2 - \frac{2}{r_1} - \frac{2}{r_2} - \nabla_{\mathbf{r}}^2 + \frac{1}{r}\,. \tag{15}$$

If we expand the two-electron wavefunction in a Taylor series in r about $r = 0$, on the (correct for the singlet state) assumption that angular terms in $\mathbf{r}$ can be ignored at low order,

$$\Psi = a_0 + a_1 r + a_2 r^2 + \ldots \tag{16}$$

then the Schrödinger equation expands as

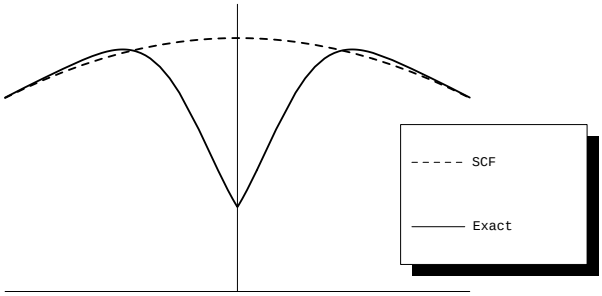
$$0 = r^{-1}\left(a_0 - 2a_1\right) + r^0\left(a_1 - 6a_2 - 4R^{-1} - E\right) + r^1\left(\ldots\right. \tag{17}$$

The r^{-1} singularity is removed if $a_1 = \frac{1}{2}a_0$, or

$$\left.\frac{\partial\Psi}{\partial r}\right|_{r=0} = \left.\frac{\Psi}{2}\right|_{r=0}\,. \tag{18}$$

This is the well-known cusp condition[4,5,6,7], which shows that in whatever direction one moves from $r = 0$, the wavefunction increases linearly. The exact wavefunction must have the shape depicted in Figure 3, showing the existence of a *Coulomb Hole* around the point of coalescence. In Figure 3, the wavefunctions are plotted against $z = z_2 - z_1$, with the two electrons having identical x, y coordinates.

Figure 3. The interelectronic cusp



The Hartree-Fock wavefunction is

$$\Psi_{\rm RHF} = \hat{\mathcal{A}} 1s^\alpha 1s^\beta = 1s(r_1)1s(r_2)\,\tfrac{1}{\sqrt{2}}(\alpha(1)\beta(2) - \beta(1)\alpha(2)) \tag{19}$$

which has no special behaviour near coalescence; in fact it is easy to show that $\partial\Psi_{\rm RHF}/\partial r = 0$ at $r = 0$. Thus the RHF wavefunction must have the shape shown in Figure 3; clearly, it overestimates the probability of finding the two electrons close together, and this in turn implies an overestimate of the electron repulsion energy. This is consistent with the variational principle, which requires the RHF energy to be higher than the exact energy. We define the correlation energy to be

$$\mathcal{E} = E^{\rm RHF} - E^{\rm exact} \tag{20}$$

where $E^{\rm exact}$ is the lowest exact eigenvalue of Schrödinger's equation. For He, $\mathcal{E} \simeq 0.042\,\text{hartree} = 1.1\,\text{eV}$.

The above analysis for the helium ground state, consisting of two electrons with opposing spin, needs to be modified when spins are instead aligned. A triplet spin wavefunction, e.g., $\alpha(1)\alpha(2)$ is symmetric with respect to electron label exchange, and so, by the Pauli principle, the spatial wavefunction must be antisymmetric. This has the consequences that, in a picture like Figure 3, the triplet wavefunction must pass through the origin, and has dipole rather than monopole $\mathbf{r}$ angular variation. There is a corresponding cusp condition specifying $\partial^2\Psi/\partial r^2$ in terms of $\partial\Psi/\partial r$ at the coalescence point[5], but the important thing is that in the energetically important region, the electrons are already kept apart by the Pauli principle, even in Hartree-Fock, and the effects of electron correlation neglect are fairly minor. Electron correlation effects are most important for electrons with opposing spins.

A further observation for polyelectronic systems is that the biggest contributions will come from pairs of electrons which occupy the same regions of physical space. If orbitals are well localized, there will be a large contribution to the correlation energy from each doubly occupied orbital, with smaller additions from pairs consisting of two different orbitals. This leads to a rough rule of thumb, that each doubly occupied orbital contributes approximately 1 eV to the total electron correlation energy.

In atomic and molecular systems, an alternative and equivalent way of visualising two-electron correlations relative to the nuclear positions is possible. If one electron is far from the nucleus of an atom, then the second electron will prefer to

be closer to the nucleus than its Hartree-Fock average; this is termed *radial* correlation. If a first electron is, say, to the right of a nucleus, then another electron will tend to visit regions of space to the left of that nucleus more than predicted by HF; this is termed *angular* correlation.

These short-range correlation effects arising from the Coulomb hole can be represented using CI wavefunctions just as with the long-range correlations discussed above. The simplest such wavefunction representing the angular correlation in the helium atom would have the form

$$\Psi = \hat{\mathcal{A}}\left(1s^{\alpha}(1)1s^{\beta}(2) + \lambda\left(2p_x^{\alpha}(1)2p_x^{\beta}(2) + 2p_y^{\alpha}(1)2p_y^{\beta}(2) + 2p_z^{\alpha}(1)2p_z^{\beta}(2)\right)\right) \quad (21)$$

It is straightforward to show that such an ansatz introduces explicit r_{12} dependence into the wavefunction. This demonstrates that CI does support correlated wavefunctions. However, unfortunately, the r_{12} dependence introduced is entirely in terms of r_{12}^2; there are no linear terms. A CI wavefunction can *never* satisfy the cusp condition (18), since its gradient will always be zero at coalescence; however, given sufficient terms, the linear combination of functions of r_{12}^2 will give a reasonable representation of the shape of the Coulomb hole. Because the expansion functions are not ideally suited to the problem, the convergence of the CI expansion is unfortunately slow, and this is discussed further below.

Historically, even some of the earliest molecular electronic structure calculations[8,9] used 2-electron basis functions of a type better adapted to the problem than orbital products (i.e., CI). Inclusion of linear terms in r_{12} is an efficient way to obtain an accurate wavefunction with a small number of functions, and probably it will remain the approach of choice when very high accuracy is needed, particularly for atoms. However, despite successful research activity in this area[10,11] this approach has not yet emerged as the best method generally applicable to molecules; CI expansions remain computationally preferable. The reason for this preference is that, although very large numbers of basis functions might be required, the hamiltonian integrals which have to be computed for CI are much simpler than for explicitly correlated wavefunctions. The explicit r_{12} terms introduce 3- and 4-electron integrals[12,13] which are potentially very numerous. In contrast, CI needs only the two-electron integrals required in an SCF calculation. Although the 3- and 4-electron integrals can be reasonably approximated[11], explicitly correlated wavefunctions still remain a specialist rather than general-purpose tool.

1.4 Second Quantization

The adoption of the CI (or other related) approach to electron correlation implies that we deal with wavefunctions which are represented as vectors in a linear space of Slater determinants; this space is in turn a subspace of N-fold products of orbitals. For the moment, we will assume that we generate all of the N-electron basis functions that we can after appropriate symmetry adaptation (electron antisymmetry, point group, etc.). Therefore the N-electron basis set is determined entirely by a choice of 1-electron basis. Before considering what this choice should be for optimum accuracy, we consider the analysis and manipulation of N-electron functions of this orbital-product type. We note initially that the orbital basis will contain at least the SCF occupied orbitals, denoted $\{\phi_i\}$, but in order that further

configurations be generated, it must be augmented by *virtual* or *external* orbitals, $\{\phi_a\}$. Both the occupied and virtual orbitals can be considered as linear combinations of an underlying chosen fixed basis $\{\chi_\alpha\}$, which will usually be atom-centred functions, exactly as in basis-set SCF calculations. The functions ϕ_p and χ_α depend only on the spatial coordinate $\mathbf{r}$; where spin-orbitals are required, they will be denoted by $\psi_p(\mathbf{x})$ and can be constructed as a product of a spatial orbital $\phi_p(\mathbf{r})$ and a spin function α or β.

Consider a complete (infinite) one particle basis set $\{\phi_p(\mathbf{r}), p = 1, 2, \dots\}$; any function of the position $\mathbf{r}$ can be represented as a linear combination of the spatial orbitals

$$f(\mathbf{r}) = \sum_p x_p \phi_p(\mathbf{r}) \; . \tag{22}$$

For a system of N electrons, a complete spatial basis can then be generated by taking all possible products $\phi_{p_1}(\mathbf{r}_1)\phi_{p_2}(\mathbf{r}_2)\dots\phi_{p_N}(\mathbf{r}_N)$, i.e., any N particle spatial function may be expanded as

$$F(\mathbf{r}_1, \mathbf{r}_2, \dots \mathbf{r}_N) = \sum_{p_1 p_2 \dots p_N} X_{p_1 p_2 \dots p_N} \phi_{p_1}(\mathbf{r}_1)\phi_{p_2}(\mathbf{r}_2)\dots\phi_{p_N}(\mathbf{r}_N) \; . \tag{23}$$

This fact is not much use for practical calculations, since we cannot use an infinite set of functions, but if we consider now the case of a finite one particle basis $\{\phi_p, p = 1, 2, \dots, m\}$, then we see the concept of the corresponding complete N particle space, composed of all possible products of orbitals. A variational calculation in such a basis will yield the lowest possible energy eigenvalue for the given one particle basis set, and such a calculation is termed Full or Complete configuration interaction (FCI). It is, however, easily appreciated that the number of possible orbital products m^N (m one electron α and β spin orbitals, N electrons) can become exceedingly large.

We introduce the useful concept of *second quantization* by defining the *orbital excitation operator* as (assuming orthogonal orbitals)

$$\hat{E}_{pq} = \sum_{i=1}^{N} |\phi_p(i)\rangle\langle\phi_q(i)| \; . \tag{24}$$

The Dirac bracket notation means that whenever the brackets become closed, $\langle f(i)|g(i)\rangle$, integration over the coordinates of electron i is performed on the functions within the bracket, $\int d\tau_i f^*(i)g(i)$. If $\hat{E}_{pq}$ is made to act on any N electron function which is a product of orbitals, or a linear combination of such products, the effect is for each occurrence of ϕ_q to generate a function which is identical, but with ϕ_q replaced by ϕ_p. Thus if ϕ_q does not appear, $\hat{E}_{pq}$ annihilates the function.

$\hat{E}_{pq}$ is a spatial orbital excitation operator; it acts on space coordinates and does not affect spin. In fact, it can be decomposed into a sum of operators which excite α and β spin orbitals separately, $\hat{E}_{pq} = \hat{e}^{\alpha}_{pq} + \hat{e}^{\beta}_{pq} = \hat{\eta}^{\alpha\dagger}_p \hat{\eta}^{\alpha}_q + \hat{\eta}^{\beta\dagger}_p \hat{\eta}^{\beta}_q$, where $\hat{\eta}^{\alpha}_q$ destroys α spin orbital ψ_q and $\hat{\eta}^{\alpha\dagger}_p$ creates α spin orbital ψ_p. The idea of second quantization is that the orbitals themselves now become quantum mechanical operators. Thus a Slater determinant can be viewed as arising from successive applications of creation

operators on the empty (vacuum) state,

$$\ldots \eta_r^\dagger \eta_q^\dagger \eta_p^\dagger \Psi_{\text{vacuum}} = \hat{\mathcal{A}}(\psi_p \psi_q \psi_r \ldots) \,. \tag{25}$$

The analysis that follows continues to use pure spatial orbitals ϕ_p; however, exactly analogous results are obtained by using explicit spin-orbitals ψ_p and spin-orbital excitation operators $\hat{e}_{pq}$. Further details of the properties of the second quantization can be found in the literature [14].

As well as the single orbital excitation operators $\hat{E}_{pq}$, it is possible to define multiple excitation operators:

$$\hat{E}_{pq,rs} = \sum_{i \neq j}^{N} |\phi_p(i)\rangle\langle\phi_q(i)| \; |\phi_r(j)\rangle\langle\phi_s(j)| \equiv \hat{E}_{rs,pq} \tag{26}$$

$$\hat{E}_{pq,rs,tu} = \sum_{i \neq j \neq k}^{N} |\phi_p(i)\rangle\langle\phi_q(i)| \; |\phi_r(j)\rangle\langle\phi_s(j)| \; |\phi_t(k)\rangle\langle\phi_u(k)| \tag{27}$$

etc.

These can all be formulated as combinations of the single excitations:

$$\hat{E}_{pq,rs} = \sum_{i,j}^{N} |\phi_p(i)\rangle\langle\phi_q(i)| \; |\phi_r(j)\rangle\langle\phi_s(j)| - \sum_{i}^{N} |\phi_p(i)\rangle\langle\phi_q(i)|\phi_r(i)\rangle\langle\phi_s(i)| \tag{28}$$

$$= \hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps} \tag{29}$$

Similar consideration of the identical operator $\hat{E}_{rs,pq}$ yields the commutation relation for the single excitations:

$$[\hat{E}_{pq}, \hat{E}_{rs}] = \hat{E}_{pq}\hat{E}_{rs} - \hat{E}_{rs}\hat{E}_{pq} = \delta_{qr}\hat{E}_{ps} - \delta_{ps}\hat{E}_{rq} \,. \tag{30}$$

Given that any wavefunction Ψ we construct is ultimately composed as a linear combination in the space of orbital products, then the following completeness identity is true for all $i = 1, 2, \ldots, N$

$$\left(\sum_{p}^{m} |\phi_p(i)\rangle\langle\phi_p(i)| \right) \; |\Psi\rangle = |\Psi\rangle \,. \tag{31}$$

Now we insert this identity into the electronic hamiltonian operator

$$\hat{H} = Z + \sum_{i}^{N} \hat{h}(i) + \sum_{i>j}^{N} r_{ij}^{-1} \,, \tag{32}$$

where Z is the nuclear repulsion energy, r_{ij} are the separations of the electrons, and $\hat{h}(i)$ is the single particle hamiltonian for each electron, incorporating its kinetic energy and the field of all the nuclei. This has the effect of replacing $\hat{H}$ by the effective *model* or *second quantized hamiltonian* $\hat{H}_M$, with the understanding that

the only thing we will ever do with $\hat{H}_M$ is to take matrix elements between functions in the orbital product space:

$$\begin{aligned}\hat{H}_M = Z &+ \sum_{i}^{N}\sum_{pq}^{m} |\phi_p(i)\rangle \langle\phi_p(i)|\hat{h}(i)|\phi_q(i)\rangle \langle\phi_q(i)| \\ &+ \sum_{i>j}^{N}\sum_{pqrs}^{m} |\phi_p(i)\rangle|\phi_r(j)\rangle \langle\phi_p(i)|\langle\phi_r(j)|r_{ij}^{-1}|\phi_q(i)\rangle|\phi_s(j)\rangle \langle\phi_q(i)|\langle\phi_s(j)| \qquad (33)\end{aligned}$$

$$= Z + \sum_{pq} h_{pq}\hat{E}_{pq} + \tfrac{1}{2}\sum_{pqrs}(pq|rs)\hat{E}_{pq,rs} \ , \qquad (34)$$

where we introduce the one and two electron *hamiltonian integrals*

$$h_{pq} = \langle\phi_p|\hat{h}|\phi_q\rangle = \int d\mathbf{r}_1 \, \phi_p^*(1)\hat{h}(1)\phi_q(1) \qquad (35)$$

$$\begin{aligned}(pq|rs) &= \langle\phi_p(1)|\langle\phi_r(2)|r_{12}^{-1}|\phi_q(1)\rangle|\phi_s(2)\rangle \\ &= \int d\mathbf{r}_1 \int d\mathbf{r}_2 \, \phi_p^*(1)\phi_r^*(2)r_{12}^{-1}\phi_q(1)\phi_s(2) \ . \qquad (36)\end{aligned}$$

For matrix elements between the N electron basis functions we then have

$$\begin{aligned}\langle\Phi_I|\hat{H}|\Phi_J\rangle &= \langle\Phi_I|\hat{H}_M|\Phi_J\rangle \\ &= Z\langle\Phi_I|\Phi_J\rangle + \sum_{pq} h_{pq}\langle\Phi_I|\hat{E}_{pq}|\Phi_J\rangle + \tfrac{1}{2}\sum_{pqrs}(pq|rs)\langle\Phi_I|\hat{E}_{pq,rs}|\Phi_J\rangle \ . \qquad (37)\end{aligned}$$

In this way, we separate *integrals* $h_{pq}, (pq|rs)$ and *coupling coefficients* $d_{pq}^{IJ} = \langle\Phi_I|\hat{E}_{pq}|\Phi_J\rangle$, $D_{pqrs}^{IJ} = \langle\Phi_I|\hat{E}_{pq,rs}|\Phi_J\rangle$. The coupling coefficients depend only on the algebraic structure of the N electron functions, and not on such factors as molecular geometry, external fields, etc.

We illustrate the use of the second-quantized formalism by considering CI wavefunctions for two electrons. Unnormalized spin-adapted basis functions can be constructed as

$$\Phi_{\pm}^{pq} = \tfrac{1}{2}\left(\hat{\mathcal{A}}(\phi_p^\alpha\phi_q^\beta) \pm \hat{\mathcal{A}}(\phi_q^\alpha\phi_p^\beta)\right) \ , \qquad (38)$$

with the upper (+) sign for spin $S = 0$ (singlet) and the lower (−) for $S = 1$ (triplet). The total wavefunction can then be expanded in this basis as

$$\begin{aligned}\Psi &= \sum_{p\geq q} C_{pq}(1 \pm \delta_{pq})\Phi_{\pm}^{pq} \\ &= \sum_{pq} C_{pq}\Phi_{\pm}^{pq} \ , \qquad (39)\end{aligned}$$

The orbital excitation operator $\hat{E}_{rs}$ when acting on $\Phi_{\pm}^{pq}$ will completely annihilate the function if s is not equal to at least one of p, q; otherwise, each occurence of ϕ_s is replaced by ϕ_r. Thus

$$\hat{E}_{rs}\Phi_{\pm}^{pq} = (1 \pm \tau_{pq})\delta_{sq}\Phi_{\pm}^{pr} \qquad (40)$$

and then

$$\hat{E}_{rs,tu}\Phi^{pq}_{\pm} = (1 \pm \tau_{pq})\delta_{sp}\delta_{uq}\Phi^{rt}_{\pm} \ , \tag{41}$$

where τ_{pq} has the effect of swapping the labels p, q in whatever follows it. Then the action of the hamiltonian operator is

$$\hat{H}\Phi^{pq}_{\pm} = (1 \pm \tau_{pq}) \left(\sum_r h_{rq}\Phi^{pr}_{\pm} + \tfrac{1}{2}\sum_{rs}(rp|qs)\Phi^{rs}_{\pm} \right) \ , \tag{42}$$

i.e.,

$$\begin{aligned} \hat{H}\Psi &= \sum_{pq} C_{pq}\hat{H}\Phi^{pq}_{\pm} \\ &= \sum_{rs} \Phi^{rs}_{\pm} \left(K(\mathbf{C})_{rs} + 2(\mathbf{hC})_{rs} \right) \ . \end{aligned} \tag{43}$$

Here, we have defined a generalized exchange matrix $\mathbf{K}(\mathbf{C})$, which for any given coefficient matrix $\mathbf{C}$ is

$$K(\mathbf{C})_{rs} = \sum_{pq} C_{pq}(rp|qs) \ . \tag{44}$$

1.5 Orbital basis sets

Calculations with complete (infinite) orbital basis sets are impossible; therefore, one immediately wants to know how to choose optimally a finite basis set such that the CI wavefunction is as close to the exact wavefunction as possible for a given number of orbitals. Insight into this problem can be gained from the two-electron example developed above. Consider the one-electron *density matrix* generated by the wavefunction, defined as

$$d_{pq} = \langle \Psi | \hat{E}_{pq} | \Psi \rangle \tag{45}$$

For the two-electron example, it is straightforward to show using (40) that

$$d_{pq} = 2\sum_s C_{sp}C_{sq} \ , \tag{46}$$

or $\mathbf{d} = 2\mathbf{C}^\dagger\mathbf{C}$.

Suppose that we now consider truncating the basis set by deleting the last (m-th) orbital to leave $m-1$ remaining functions. The overlap between the new and old wavefunctions is

$$\begin{aligned} \langle \Psi_{\rm New} | \Psi_{\rm Old} \rangle &= \langle \Psi_{\rm Old} | \Psi_{\rm Old} \rangle - 2\sum_{pqr} C_{pq}C_{rm}\langle \Phi_{pq} | \Phi_{rm} \rangle + \sum_{pq} C_{pq}C_{mm}\langle \Phi_{pq} | \phi_{mm} \rangle \\ &= 1 - 2(\mathbf{C}^\dagger\mathbf{C})_{mm} + C^2_{mm} \end{aligned} \tag{47}$$

Ignoring the last (C^2_{mm}) term, which can be shown to be of lesser importance, we deduce that the amount that the overlap differs from unity is d_{mm}. Consider making linear transformations amongst the underlying orbitals. Of all the possible transformations, the one which minimises d_{mm} is that which brings $\mathbf{d}$ to diagonal form, with d_{mm} being the smallest eigenvalue. Such orbitals are known as *natural*

orbitals (NOs), and are of great utility in interpreting correlated many-electron wavefunctions. The trace of the density matrix is equal to the number of electrons, leading to an interpretation of the eigenvalues as occupation numbers.

In the above example, therefore, if natural orbitals are chosen, the effects of deleting the last (m-th) orbital are minimized. In other words, the CI wavefunction in $m-1$ orbitals is as good as it can be. We have thus shown that of all the possible choices of orbitals, natural orbitals offer the most compact or efficient basis set, for a two-electron system. For many-electron systems, the situation is, of course, more complicated. One can still define natural orbitals as density matrix eigenvalues, but their relationship with the wavefunction is not so transparent. For the special case of CI wavefunctions that contain up to double excitations from the Hartree-Fock determinant, then one can also construct *pair natural orbitals* (PNOs) for each pair of occupied orbitals that are excited; these PNOs do have similar properties to the two-electron NOs, and typically show a similar convergence of eigenvalues towards zero. The true NOs, however, are an average of the various PNOs, and the convergence of their spectrum and their usefulness in evaluating the correlating effect of basis functions is usually less advantageous.

In contrast to Hartree-Fock, where reasonably good wavefunctions can be obtained using a double-zeta plus polarization (DZP) basis set allowing for simple contraction and deformation of atomic orbitals, a much larger basis set is required for recovering a large fraction of the correlation energy; i.e., the sequence of NO occupation numbers is found to be rather slowly convergent. It is then not a trivial problem to decide straightaway what basis functions $\{\chi_\alpha\}$ should be used for optimum recovery of electron correlation effects. The idea of using natural orbitals to obtain basis sets is taken to the extreme in the atomic natural orbital (ANO) basis scheme[15]. Here, the basis functions are (approximate) *atomic* natural orbitals, obtained from a CI calculation on each of the molecule's constituent atoms. The idea is that the ANOs, which are near–optimum correlating functions for the atomic problem, will be good functions for describing molecular electron correlation. Within each of the atomic symmetries $(s, p, d, \dots)$, each contracted basis function is a linear combination of all the primitive gaussian functions; thus each primitive function enters in to all contractions (*general contraction*). Within the ANO scheme, there also arises the concept of sequences of basis sets, in which each basis set is derived from the previous one by the addition of the next most important atomic natural orbital. This allows for the systematic improvement of basis sets and consequent elimination of possible spurious errors arising from unbalanced choices of basis functions. For example, for most first row atoms, examination of the ANO occupation numbers identifies $[3s2p1d]$, $[4s3p2d1f]$ and $[5s4p3d2f1g]$ as good choices of contracted basis sets, whilst a set such as $[5s3p2d2f2g]$ is unbalanced, and would be inefficient in recovering electron correlation effects.

For certain applications, selection of a small or medium–sized ANO set will not necessarily result in a good basis set, and can lead to spurious results. An example is the calculation of atomic or molecular electrical polarizabilities. Here, it is vital to include diffuse basis functions, particularly of d type in the usual case that the highest atomic shell is of p type. Such basis functions do not appear in the set which is optimum for the correlation problem, and so such functions must

be included additionally, or the basis set redesigned somewhat. This case occurs to a milder degree in all molecules, where the atomic functions are polarized by their neighbours; even for SCF calculations, polarization functions are required to cover this effect, and the optimum gaussian exponents are not necessarily related to those best for correlation. Another type of calculation which presents problems for ANO sets is that where several different atomic states are involved; the classic case is in transition metal chemistry, where $d^n s^2$, $d^{n+1} s^1$ and d^{n+2} atomic states often all make significant contributions to the molecular situation. ANO bases based on each state are drastically different, particularly for the d orbitals, which are much more diffuse in d^{n+2} than in d^n; so the use of an ANO set derived from one particular atomic state can introduce an unwanted bias towards that state. A partial solution is to select functions which are eigenfunctions of the sum of the density matrices for each state[16,17,18], although caution is still needed. For general applications, a good compromise is found in the "correlation consistent" basis sets[19], which are similar to ANO sets, except that the most diffuse s and p functions are left uncontracted, and the polarization functions are simple uncontracted gaussians designed to cover both the polarization and correlation requirements. In fact, the advantage in using ANOs for the polarization functions is not that great, and the correlation consistent basis sets are usually more compact than standard ANOs for a given level of accuracy. Just as with ANOs, a systematic sequence of basis sets is defined, with members conventionally denoted cc-pVDZ, cc-pVTZ, cc-pVQZ, cc-pV5Z, etc., which for 1st row atoms comprise $3s2p1d$, $4s3p2d1f$, $5s4p3d2f1g$, $6s5p4d3f2g1h\ldots$.

1.6 Dynamical vs. Non-Dynamical Correlation

The correlation energy arising from overestimation of short-range electron repulsions in Hartree-Fock wavefunctions is usually referred to as dynamical correlation. Dynamical correlation is always reduced when a normal chemical bond (i.e., doubly occupied orbital) is broken. It is the neglect of dynamical correlation which causes the RHF equilibrium energy of F_2 to be higher than twice the RHF energy of a fluorine atom, since in F_2 there are 9 pairs of electrons, but in each F there are only 4. The effect is so pronounced for F because the molecular orbitals are considerably smaller than their atomic parents, and crowding the electrons together means there is more correlation energy. Where dynamical correlation effects are important, Hartree-Fock will therefore generally overestimate bond lengths and underestimate binding. An extreme example is that of rare-gas dimers, which are unbound at the Hartree-Fock level, but in reality are held together by dispersion, which is a manifestation of dynamic correlation.

That part of the correlation energy arising from long-range correlation effects, such as observed on molecular dissociation, is often referred to as non-dynamical (or static) correlation. Static correlation effects mean that (spin-restricted) Hartree-Fock tends to artificially overbind molecules underestimating bond lengths and overestimating vibrational frequencies. Thus the effects of dynamic and non-dynamic correlation are very often in opposition, and the partial cancellation of correlation errors enhances the value of SCF; it is often observed that, for example, use of methods which represent properly the non-dynamical correlation effects leads to

much worse agreement of computed properties with experiment than RHF.

The division between dynamical and non-dynamical correlation is difficult to define in most cases. For example, when thinking about electron correlation in a bond in a molecule, the radial and angular short-range concepts are somewhat blurred with the ideas of long-range dissociation-enabling correlation. One useful visualization is that the non-dynamical correlation is that which is recovered with the minimum CI expansion describing properly all correlation effects; in contrast, convergence of the dynamical correlation energy with increasing size of CI expansion is very slow.

When non-dynamical correlation is weak, Hartree-Fock theory already provides a qualitatively correct description of the wavefunction. Under such circumstances, which, fortunately, apply for the majority of molecules in their ground state near equilibrium geometry, one may use *single-reference* methods for representing the dynamical correlation effect. These methods build on the SCF reference determinant, typically using perturbative arguments to define classes of configurations or excitations deemed to be of most importance in constructing an approximate correlated wavefunction. For most excited states, for molecules that are close to dissociation, and for situations in which there is near electronic degeneracy, Hartree-Fock is a poor approximation. Static correlation effects often mean that there is no single Slater determinant that dominates the wavefunction, and perturbative or other approaches that assume a good single-reference starting point are doomed to failure. Under such circumstances, a viable way forward is to first deal with the static correlation problem using a CI expansion that covers all the important effects. One may then go further using this many-determinant reference as a starting point for further recovery of the dynamic correlation. Such approaches are termed *multi-reference* methods.

2 Closed-Shell single-reference methods

In this section we will discuss the most important electron correlation methods based on closed-shell Hartree-Fock reference functions. This includes Møller-Plesset perturbation theory, singles and doubles configuration interaction (CISD), and non-variational variants like the coupled-electron pair approximation (CEPA), as well as coupled cluster methods with single and double excitations (CCSD). The effect of triple excitations can be accounted for by perturbation theory, leading to CCSD(T).

From a computational point of view, it is important to minimize the logic in the code, and to formulate the theory in terms of matrix and vector operations. The most efficient operations one can perform on any kind of current hardware are matrix multiplications. This applies both to vector computers as well as to RISC workstations or even PCs. The reason for this is that on most machines the bottleneck is not the floating point operation itself, but getting the data from the memory, in particular if the quantities involved do not fit into the fast cache. By an appropriate unrolling of the three loops in a matrix multiplication one can achieve that each data element obtained from memory can be used in several floating point operations, and this this way often about 80% of the theoretical peak performance can be achieved.

For the formulation of the theory in terms of matrix multiplications it is essential to use unnormalized or even non-orthogonal configuration state functions. We start with a general discussion of the configuration spaces which are common to all methods discussed in the subsequent sections.

2.1 The first-order interacting space

According to second-order perturbation theory, the most important contributions to the correlation energy arise from configurations Φ_I which have non-zero matrix elements $\langle\Phi_I|\hat{H}|\Phi^{SCF}\rangle$, i.e., which span the *first-order interacting space* of Φ^{SCF}. In the following, the SCF wavefunction will be denoted $|0\rangle \equiv \Phi^{SCF}$. According to the Slater-Condon rules only Slater determinants can contribute which differ by at most two spin-orbitals from the Hartree-Fock determinant. The spin adapted singly and doubly excited configurations are conveniently generated by applying the excitation operators $\hat{E}_{ai}$ to the reference function

$$\Phi_i^a = \hat{E}_{ai}|0\rangle \ , \tag{48}$$

$$\Phi_{ij}^{ab} = \hat{E}_{ai}\hat{E}_{bj}|0\rangle \ , \tag{49}$$

where i, j refer to occupied orbitals in $|0\rangle$, and a, b to virtual orbitals (unoccupied in $|0\rangle$). If $|0\rangle$ is an optimized closed-shell Hartree-Fock wavefunction, the matrix elements $\langle\Phi_i^a|\hat{H}|0\rangle = 2f_{ai}$ vanish for all single replacements Φ_i^a, since the optimized orbitals satisfy the conditions $f_{ai} = 0$ (Brillouin theorem). Therefore, the first-order wavefunction is a linear combination of all doubly excited configurations Φ_{ij}^{ab}

$$\Psi^{(1)} = \frac{1}{2}\sum_{ij}\sum_{ab} T_{ab}^{ij}\Phi_{ij}^{ab} \ , \tag{50}$$

where T_{ab}^{ij} are the amplitudes. Note that the operators $\hat{E}_{ai}$ and $\hat{E}_{bj}$ commute, and therefore

$$\Phi_{ij}^{ab} = \Phi_{ji}^{ba} \ , \tag{51}$$

i.e., the configuration set used in the expansion of $\Psi^{(1)}$ is redundant. In the formulation of correlation theories it will be convenient to use this redundant set, but we must account for this by the restriction

$$T_{ab}^{ij} = T_{ba}^{ji} \ . \tag{52}$$

We will consider T_{ab}^{ij} as matrices with elements ab. Different matrices are labeled by the superscripts ij:

$$[\mathbf{T}^{ij}]_{ab} = T_{ab}^{ij} \ , \qquad \mathbf{T}^{ij} \ = \ \mathbf{T}^{ji\dagger} \ . \tag{53}$$

The matrix elements for $i > j$, all a, b and $i = j$, $a \geq b$ form the non-redundant set of amplitudes.

The definition of the doubly excited configurations in eq. (49) is most simple, but has the disadvantage that the resulting functions are non-orthogonal. Using the commutation relations (30) and the fact that zero results if an external annihilator acts on the reference function $|0\rangle$ one obtains

$$\langle\Phi_{ij}^{ab}|\Phi_{kl}^{cd}\rangle = \delta_{ac}\delta_{bd}\langle 0|\hat{E}_{ik,jl}|0\rangle + \delta_{ad}\delta_{bc}\langle 0|\hat{E}_{il,jk}|0\rangle \ , \tag{54}$$

where $\langle 0|\hat{E}_{ik,jl}|0\rangle$ are the elements of the second-order reduced density matrix of the reference function. For closed-shell Hartree-Fock reference functions one obtains explicitly

$$\begin{aligned}\langle 0|\hat{E}_{ik,jl}|0\rangle &= 4\delta_{ik}\delta_{jl} - 2\delta_{il}\delta_{jk} \ , \\ \langle \Phi_{ij}^{ab}|\Phi_{kl}^{cd}\rangle &= \delta_{ac}\delta_{bd}(4\delta_{ik}\delta_{jl} - 2\delta_{il}\delta_{jk}) + \delta_{ad}\delta_{bc}(4\delta_{il}\delta_{jk} - 2\delta_{ik}\delta_{jl}) \ . \end{aligned} \tag{55}$$

Straightforward use of these non-orthogonal configurations is in principle possible, but leads to some complications. There are two ways for simplification: in the first case a set of orthogonal configuration state functions is defined as

$$\Phi_{ijp}^{ab} = \frac{1}{2}\left(\Phi_{ij}^{ab} + p\Phi_{ij}^{ba}\right) \qquad \text{for } p = \pm 1,\ i \geq j,\ a \geq b \ , \tag{56}$$

where $p = 1$ corresponds to singlet coupling of the two external electrons, and $p = -1$ to triplet coupling. Note that these functions are not normalized; for a closed-shell reference function we have

$$\langle \Phi_{ijp}^{ab}|\Phi_{klq}^{cd}\rangle = (2-p)\delta_{pq}(\delta_{ac}\delta_{bd} + p\delta_{ad}\delta_{bc})(\delta_{ik}\delta_{jl} + p\delta_{il}\delta_{jk}) \ , \tag{57}$$

and thus the normalization factors are

$$\langle \Phi_{ijp}^{ab}|\Phi_{ijq}^{ab}\rangle = (2-p)\delta_{pq}(1 + p\delta_{ab})(1 + p\delta_{ij}) \ . \tag{58}$$

As will become clear later, for an efficient formulation of all electron correlation methods it is essential not to normalize the configurations. This was first realized in the theory of self-consistent electron pairs (SCEP) by Meyer[20], who showed that by using unnormalized configurations all terms involving the virtual orbital labels $a, b, \ldots$ can be formulated in a computationally convenient matrix form without any logic. Most importantly, this concerns the factor $(1+p\delta_{ab})$, which implies a different normalization for *diagonal* configurations ($a = b$) than for non diagonal ones ($a \neq b$). We note that in the original SCEP theory of Meyer[20] the configurations were normalized by the factors $[(2-p)(1+p\delta_{ij})]^{-1/2}$, but this leads to some unnecessary factors in the resulting equations. A similar definition is possible for multireference wavefunctions and will be used in section 5.

For single-reference methods it turns out that even simpler equations can be obtained by directly using the configurations (49) together with a set of *contravariant* configurations[21,22]

$$\tilde{\Phi}_{ij}^{ab} = \frac{1}{6}(2\Phi_{ij}^{ab} + \Phi_{ji}^{ab}) \tag{59}$$

which have the properties

$$\langle \tilde{\Phi}_{ij}^{ab}|\Phi_{kl}^{cd}\rangle = \delta_{ac}\delta_{bd}\delta_{ik}\delta_{jl} + \delta_{ad}\delta_{bc}\delta_{il}\delta_{jk} \ , \tag{60}$$

$$\langle \tilde{\Phi}_{ij}^{ab}|\Psi^{(1)}\rangle = T_{ab}^{ij} \ , \tag{61}$$

$$\langle \tilde{\Phi}_{ij}^{ab}|\hat{H}|\Psi^{(0)}\rangle = (ai|bj) \ . \tag{62}$$

The last expression is obtained by inserting the hamiltonian in second quantization (cf. eq. (34))

$$\langle \tilde{\Phi}_{ij}^{ab}|\hat{H}|\Psi^{(0)}\rangle = \frac{1}{2}\sum_{rstu}\langle \tilde{\Phi}_{ij}^{ab}|\hat{E}_{rs,tu}|\Psi^{(0)}\rangle(rs|tu) \ , \tag{63}$$

and realizing that the indices r, t must be external and match a, b, while s, u must be internal and match i, j according to eq. (60)

$$\begin{aligned}\langle\tilde{\Phi}_{ij}^{ab}|\hat{H}|\Psi^{(0)}\rangle &= \frac{1}{2}\sum_{kl}\sum_{cd}\langle\tilde{\Phi}_{ij}^{ab}|\hat{E}_{ck}\hat{E}_{dl}|\Psi^{(0)}\rangle(ck|dl) \\ &= \frac{1}{2}\sum_{kl}\sum_{cd}\langle\tilde{\Phi}_{ij}^{ab}|\Phi_{kl}^{cd}\rangle(ck|dl) \;=\; (ai|bj)\;. \end{aligned} \tag{64}$$

We can now express $\Psi^{(1)}$ either in the original basis or in the basis of contravariant functions

$$\Psi^{(1)} = \frac{1}{2}\sum_{ij}\sum_{ab} T_{ab}^{ij}\Phi_{ij}^{ab} \;=\; \sum_{ij}\sum_{ab}\tilde{T}_{ab}^{ij}\tilde{\Phi}_{ij}^{ab}\;, \tag{65}$$

which leads to

$$\tilde{T}_{ab}^{ij} = 2T_{ab}^{ij} - T_{ab}^{ji} \qquad \text{or} \qquad \tilde{\mathbf{T}}^{ij} \;=\; 2\mathbf{T}^{ij} - \mathbf{T}^{ji}\;. \tag{66}$$

The factor $\frac{1}{2}$ has been omitted in the second sum for convenience in later expressions.

For the singles we can define the contravariant space analogously, but in this case only the normalization of Φ_i^a and $\tilde{\Phi}_i^a$ differs

$$\tilde{\Phi}_i^a = \frac{1}{2}\Phi_i^a\;, \tag{67}$$

$$\tilde{t}_a^i = 2t_a^i\;. \tag{68}$$

2.2 Matrix notation

We have seen above that the amplitudes T_{ab}^{ij} for a given correlated orbital pair (ij) can be considered as a matrix $\mathbf{T}^{ij}$, and the amplitudes t_a^i of the single excitations as vectors $\mathbf{t}^i$. Unless otherwise noted, here and in the following i, j, k, l refer to occupied orbitals, a, b, c, d to virtual orbitals (unoccupied in the reference function), and p, q, r, s to any orbitals. In open-shell and MCSCF methods, t, u, v, w will denote open-shell (active) orbitals.

Similarly, it is convenient to order the two-electron integrals over two occupied and two virtual orbitals into matrices. In this case there are two types, namely Coulomb and exchange matrices

$$J_{ab}^{ij} = (ab|ij)\;, \tag{69}$$

$$K_{ab}^{ij} = (ai|bj)\;. \tag{70}$$

The labels ij refer to different matrices, and ab to their elements. Often it will be possible to write equations in matrix form, involving matrix multiplications and additions, and then bold face letters will be used for matrices, e.g., $\mathbf{J}^{ij}$ and $\mathbf{K}^{ij}$. For convenience in later expressions, we also define

$$L_{ab}^{ij} = 2K_{ab}^{ij} - K_{ba}^{ij}\;, \tag{71}$$

and the closed shell Fock matrix

$$f_{rs} = h_{rs} + \sum_i \left[2J_{rs}^{ii} - K_{rs}^{ii}\right]\;. \tag{72}$$

In the subsequent sections, the matrix **f** will only refer to the external part, i.e, the elements f_{ab}.

2.3 Second-order Møller-Plesset perturbation theory

The simplest electron correlation method to treat electron correlation is Møller-Plesset perturbation theory, which is a special variant of Rayleigh-Schrödinger perturbation theory, with the zeroth-order hamiltonian

$$\hat{H}^{(0)} = \sum_{i=1}^{N_{el}} \hat{f}(i) = \sum_{rs} \hat{E}_{rs} f_{rs} \ , \tag{73}$$

and with

$$\hat{H}^{(1)} = \hat{H} - \hat{H}^{(0)} \ , \tag{74}$$

where $\hat{f}(i)$ is the closed-shell Fock operator for electron i. For optimized orbitals the matrix elements f_{ai} vanish (Brillouin conditions), and it is then easily shown that the Hartree-Fock wavefunction $\Psi^{(0)} = \Phi^{\mathrm{SCF}}$ is an eigenfunction of $\hat{H}^{(0)}$, i.e.,

$$\hat{H}^{(0)}\Psi^{(0)} = \hat{E}^{(0)}\Psi^{(0)} \ , \tag{75}$$

$$\hat{E}^{(0)} = 2\sum_{i=1}^{m_{occ}} f_{ii} \ , \tag{76}$$

$$\hat{E}^{(0)} + \hat{E}^{(1)} = \langle\Psi^{(0)}|\hat{H}|\Psi^{(0)}\rangle \ = \ E^{\mathrm{SCF}} \ , \tag{77}$$

where E^{SCF} is the Hartree-Fock energy expectation value.

The first-order wavefunction is expanded according to eq. (50), and the amplitudes T_{ab}^{ij} are obtained by solving the first-order perturbation equations

$$\langle\tilde{\Phi}_{ij}^{ab}|\hat{H}^{(0)} - \hat{E}^{(0)}|\Psi^{(1)}\rangle + \langle\tilde{\Phi}_{ij}^{ab}|\hat{H}|\Psi^{(0)}\rangle = 0 \tag{78}$$

for all $i \geq j, ab$. Inserting eq. (50) and evaluating the matrix elements yields the linear equations

$$R_{ab}^{ij} = K_{ab}^{ij} + \sum_c \left(f_{ac}T_{cb}^{ij} + T_{ac}^{ij}f_{cb} \right) - \sum_k \left(f_{ik}T_{ab}^{kj} + T_{ab}^{ik}f_{kj} \right) \ = \ 0 \ . \tag{79}$$

For the case that *canonical* Hartree-Fock orbitals are used which obey

$$f_{ij} = \epsilon_i \delta_{ij} \ , \tag{80}$$

$$f_{ab} = \epsilon_a \delta_{ab} \ , \tag{81}$$

one obtains

$$R_{ij}^{ab} = K_{ab}^{ij} + (\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j)T_{ab}^{ij} \tag{82}$$

$$T_{ab}^{ij} = -K_{ab}^{ij}/(\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j) \ , \tag{83}$$

which is, of course, the well known MP2 expression. Using eqs. (61) and (62) the second-order energy takes the form

$$\begin{aligned}\hat{E}^{(2)} &= \langle\Psi^{(0)}|\hat{H}|\Psi^{(1)}\rangle \\ &= \sum_{ij}\sum_{ab}\langle\Psi^{(0)}|\hat{H}|\tilde{\Phi}_{ij}^{ab}\rangle\tilde{T}_{ab}^{ij} \\ &= \sum_{ij}\langle\mathbf{K}^{ij}\tilde{\mathbf{T}}^{ji}\rangle \;=\; \sum_{ij}\langle\mathbf{K}^{ij}(2\mathbf{T}^{ji}-\mathbf{T}^{ij})\rangle \;, \end{aligned} \tag{84}$$

where

$$\langle\mathbf{K}^{ij}\tilde{\mathbf{T}}^{ji}\rangle = \sum_{ab} K_{ab}^{ij}\tilde{T}_{ba}^{ji} \;=\; \sum_{ab} K_{ab}^{ij}\tilde{T}_{ab}^{ij} \tag{85}$$

$$\tag{86}$$

denotes the trace of the matrix product in the brackets.

From the above equations it is obvious that evaluating the second-order energy is trivial once the exchange integrals $K_{ab}^{ij} = (ai|bj)$ are available. These integrals are in the MO basis, and must therefore be generated from the 2-electron integrals in the AO basis by a four-index transformation

$$(ai|bj) = \sum_{\mu\nu\rho\sigma} X_{\mu a}X_{\nu b}X_{\rho i}X_{\sigma j}(\mu\rho|\nu\sigma) \;. \tag{87}$$

This transformation is most efficiently done in four steps, each being a matrix multiplication, i.e.

$$(\mu\rho|\nu j) = \sum_{\sigma}(\mu\rho|\nu\sigma)X_{\sigma j} \;, \tag{88}$$

$$(\mu i|\nu j) = \sum_{\rho}(\mu\rho|\nu j)X_{\rho i} \;, \tag{89}$$

$$(\mu i|bj) = \sum_{\nu}(\mu j|\nu i)X_{\nu b} \;, \tag{90}$$

$$(ai|bj) = \sum_{\mu}(\mu i|bj)X_{\mu a} \;. \tag{91}$$

Since the number of occupied orbitals i, j is usually much smaller than the number of basis functions, the number of transformed integrals becomes smaller in each step, and therefore the first quarter transformation step is most expensive. It requires about $\frac{1}{2}m_{val}m^4$ operations, where m is the number of basis functions and m_{val} the number of correlated orbitals. Since both m_{val} and m increase linearly with system size $\mathcal{N}$, the computational effort scales with $\mathcal{O}(\mathcal{N}^5)$. For large systems not only the computation time but also the storage of the two-electron integrals and intermediate quantities is a severe bottleneck. Chapter 6 discusses *integral-direct* transformations, in which the integrals $(\mu\rho|\nu\sigma)$ are computed on the fly whenever needed, without being ever stored on disk.

An alternative way to compute the second-order energy is to start from the *Hylleraas functional*

$$\begin{aligned} E_2 &= 2\langle\Psi^{(1)}|\hat{H}|\Psi^{(0)}\rangle + \langle\Psi^{(1)}|\hat{H}^{(0)} - \hat{E}^{(0)}|\Psi^{(1)}\rangle \\ &= 2\sum_{ij}\left[\langle\mathbf{K}^{ij}\tilde{\mathbf{T}}^{ji}\rangle + \langle\mathbf{T}^{ij}\mathbf{f}\tilde{\mathbf{T}}^{ji}\rangle - f_{ij}\sum_k\langle\mathbf{T}^{ik}\tilde{\mathbf{T}}^{kj}\rangle\right] \\ &= \sum_{ij}\left[\langle(\mathbf{K}^{ij} + \mathbf{R}^{ij})\tilde{\mathbf{T}}^{ji}\rangle\right] . \end{aligned} \tag{92}$$

Minimizing this functional with respect to the $\tilde{T}_{ab}^{ij}$ yields

$$\frac{\partial E_2}{\partial \tilde{T}_{ab}^{ij}} = 2R_{ab}^{ij} , \tag{93}$$

with the $\mathbf{V}^{ij}$ defined in eq. (79). Thus, the Hylleraas functional is stationary with respect to small variations of the T_{ab}^{ij} if the first-order perturbation equations are fulfilled, i.e. $R_{ab}^{ij} = 0$. For the corresponding amplitudes we have $E_2 = \hat{E}^{(2)}$. It is straightforward to show that in general $E_2 \geq \hat{E}^{(2)}$ for any set of trial function $\Psi^{(1)}$. The stationary property is very convenient for deriving the MP2 gradient expression and in the context of local electron correlation methods to be discussed later.

Even though we will not discuss applications of the methods in this article, it should be noted that the applicability of MP2 is restricted to cases with a sufficient large HOMO-LUMO gap. If this is not the case, the energy denominators in eq. (83) become small and the perturbation expansion diverges.

2.4 Singles and doubles configuration interaction

In singles and doubles configuration interaction (CISD) the expansion coefficients are determined variationally. Consequently, the resulting energy is an upper bound to the exact energy, but it is not size extensive or size consistent, i.e., it does not scale correctly with the number of electrons or the number of independent subsystems. Therefore, CISD usually yields poor results, and it is not recommended to be used. However, much better results can be obtained by some simple modifications of the variational conditions, leading to the coupled electron pair approximation (CEPA)[23,24] or the coupled pair functional (CPF)[25], which are approximately size consistent and yield much better results at the same computational cost as CISD.

The first matrix formulation of CISD is due to Meyer and known as SCEP theory[20] (cf. section 2.1). This method was formulated originally in the AO basis, but here we will continue to work in a basis of orthogonal MOs, which is somewhat simpler. However, we will come back to the AO formulation when discussing local electron correlation theories.

The CISD wavefunction is expanded in terms of the same configurations as used in the MP2 wavefunction, but also includes single excitations

$$\Psi^{\mathrm{CISD}} = \Phi^{\mathrm{SCF}} + \sum_{ia} t_a^i \Phi_i^a + \frac{1}{2}\sum_{ij}\sum_{ab} T_{ab}^{ij}\Phi_{ij}^{ab} . \tag{94}$$

The coefficients t_a^i, T_{ab}^{ij} are optimized variationally by minimizing the Rayleigh quotient

$$E^{\mathrm{CISD}} = \frac{\langle \Psi^{\mathrm{CISD}} | \hat{H} | \Psi^{\mathrm{CISD}} \rangle}{\langle \Psi^{\mathrm{CISD}} | \Psi^{\mathrm{CISD}} \rangle} \,. \tag{95}$$

Using eqs. (61) and (65) one finds for the norm

$$\begin{aligned} N \;=\; \langle \Psi^{\mathrm{CISD}} | \Psi^{\mathrm{CISD}} \rangle &= 1 + \sum_{ai} \tilde{t}_a^i t_a^i + \sum_{ij} \sum_{ab} \tilde{T}_{ab}^{ij} T_{ab}^{ij} \\ &= 1 + \sum_i \langle \tilde{\mathbf{t}}^{i\dagger} \mathbf{t}^i \rangle + \sum_{i \geq j} (2 - \delta_{ij}) \langle \tilde{\mathbf{T}}^{ij} \mathbf{T}^{ji} \rangle \,. \end{aligned} \tag{96}$$

Differentiating the expectation value with respect to the $\tilde{T}_{ab}^{ij}$ yields the eigenvalue equations

$$\begin{aligned} r_a^i &= \langle \tilde{\Phi}_i^a | \hat{H} - E^{\mathrm{CISD}} | \Psi^{\mathrm{CISD}} \rangle \;=\; 0 \,, \\ R_{ab}^{ij} &= \langle \tilde{\Phi}_{ij}^{ab} | \hat{H} - E^{\mathrm{CISD}} | \Psi^{\mathrm{CISD}} \rangle \;=\; 0 \,. \end{aligned} \tag{97}$$

These equations can be solved iteratively (*direct CI*). In each iteration one has to compute the residuals

$$r_a^i = v_a^i - \mathcal{E}^{\mathrm{CISD}} t_a^i \,, \tag{98}$$

$$R_{ab}^{ij} = V_{ab}^{ij} - \mathcal{E}^{\mathrm{CISD}} T_{ab}^{ij} \tag{99}$$

where

$$v_a^i = \langle \tilde{\Phi}_i^a | \hat{H} - E^{\mathrm{SCF}} | \Psi^{\mathrm{CISD}} \rangle \tag{100}$$

$$V_{ab}^{ij} = \langle \tilde{\Phi}_{ij}^{ab} | \hat{H} - E^{\mathrm{SCF}} | \Psi^{\mathrm{CISD}} \rangle \,, \tag{101}$$

and $\mathcal{E}^{\mathrm{CISD}} = E^{\mathrm{CISD}} - E^{\mathrm{SCF}}$ is the correlation energy

$$\mathcal{E}^{\mathrm{CISD}} = \frac{1}{N} \left[\sum_i (f_a^i + v_a^i) \tilde{t}_a^i + \sum_{ij} \sum_{ab} (K_{ab}^{ij} + V_{ab}^{ij}) \tilde{T}_{ab}^{ij} \right] \,. \tag{102}$$

The residuals are used to obtain an update of the CI-coefficients by simple perturbation theory:

$$\Delta t_a^i = \frac{-r_a^i}{\langle \tilde{\Phi}_i^a | \hat{H} - E^{\mathrm{CISD}} | \Phi_i^a \rangle} \,, \qquad \Delta T_{ab}^{ij} \;=\; \frac{-R_{ab}^{ij}}{\langle \tilde{\Phi}_{ij}^{ab} | \hat{H} - E^{\mathrm{CISD}} | \Phi_{ij}^{ab} \rangle} \,. \tag{103}$$

This procedure relies on the fact that the hamiltonian in the configuration basis is diagonal dominant. Convergence can be improved and guaranteed by the Davidson procedure[26].

For the sake of simplicity, we will restrict the following discussion to double excitations (CID); the inclusion of single excitations is quite straightforward and does not lead to any principle difficulties. In the CID case the matrices $\mathbf{V}^{ij}$ take the explicit form

$$V_{ab}^{ij} = K_{ab}^{ij} + K(\mathbf{T}^{ij})_{ab} + \sum_{kl} K_{kl}^{ij} T_{ab}^{kl} + G_{ab}^{ij} + G_{ba}^{ji} \tag{104}$$

with the auxiliary matrices

$$\mathbf{G}^{ij} = \mathbf{T}^{ij}\mathbf{f} - \sum_k \left[\mathbf{T}^{ik} f_{kj} + \mathbf{T}^{ik}\mathbf{J}^{kj} + (\mathbf{T}^{ik}\mathbf{J}^{kj})^\dagger - \tilde{\mathbf{T}}^{ik}\mathbf{K}^{kj} \right] . \tag{105}$$

The matrices $\mathbf{G}^{ij}$ account for the contributions of the two-electron integrals over two external and two occupied orbitals, i.e., all matrices occuring in eq. (105) are defined in the space of external orbitals only. The evaluation of all $\mathbf{G}^{ij}$ requires $2m_{val}^3$ matrix multiplications. Since each matrix multiplication involves $2m_{ext}^3$ floating point operations, the total cost scales with the sixth power of the molecular size. Note the exceedingly simple matrix form of these equations, which do not involve any complicated logic. This is solely due to the fact that unormalized and non-orthogonal configurations are used, as outlined in section 2.1. In contrast, in the early direct CISD method of Roos and Siegbahn[27], which employed orthonormalized configuration state functions, about 140 different types of matrix elements had to be distinguished.

The so called *external exchange operators* $\mathbf{K}(\mathbf{T}^{ij})$ in the second term of (104) account for all contributions of integrals over four external orbitals

$$K(\mathbf{T}^{ij})_{ab} = \sum_{cd} T_{cd}^{ij}(ac|db) . \tag{106}$$

There terms require about $m_{val}^2 m_{ext}^4$ floating point operations, and for large basis sets and not too many correlated orbitals m_{val} their evaluation dominates the total computational cost. As written in eq. (106) one would need a full integral transformation for generating the integrals $(ac|dc)$. This would not only be rather expensive ($\mathcal{O}(\mathcal{N}^5)$ operations), but also double the disk space. The transformation can be avoided by expanding the virtual MOs in the integral, yielding

$$\begin{aligned} K(\mathbf{T}^{ij})_{ab} &= \sum_{\mu\nu} X_{\mu a} X_{\nu b} \sum_{\rho\sigma} \left[\sum_{cd} X_{\rho c} T_{cd}^{ij} X_{\sigma d} \right] (\mu\rho|\sigma\nu) \\ &= \sum_{\mu\nu} X_{\mu a} X_{\nu b} \sum_{\rho\sigma} T_{\rho\sigma}^{ij} (\mu\rho|\sigma\nu) \\ &= \left[\mathbf{X}^\dagger \mathbf{K}(\mathbf{T}^{ij})_{\mathrm{AO}} \mathbf{X} \right]_{ab} . \end{aligned} \tag{107}$$

The quantities $\mathbf{T}_{\mathrm{AO}}^{ij} = \mathbf{X}\mathbf{T}_{\mathrm{MO}}^{ij}\mathbf{X}^\dagger$ are the amplitudes in the AO basis. These are precomputed and then contracted with the two-electron integrals $(\mu\rho|\sigma\nu)$, which very much resembles the calculation of the exchange terms in the Fock matrix. The resulting operators in the AO basis $\mathbf{K}(\mathbf{T}^{ij})_{\mu\nu}$ are finally backtransformed into the MO basis by the two matrix multiplications in the last line. Similar operators are also needed in coupled cluster theory (cf. section 2.5) and multirefence configuration interaction (cf. section 5).

The third-order energy in Møller-Plesset perturbation energy is obtained as

$$E^{(3)} = \sum_{ij} \sum_{ab} (K_{ab}^{ij} + V_{ab}^{ij}) \tilde{T}_{ab}^{ij} , \tag{108}$$

where the V_{ab}^{ij} and $\tilde{T}_{ab}^{ij}$ are computed from the MP2 amplitudes. Note that this energy expression is similar to the expectation value, eq. (102), but without the

normalization factor. In contrast to the CID energy $E^{(3)}$ is size consistent, but not an upper bound to the exact energy.

Finally, we note that the CEPA equations[23,24] can be obtained from the CISD equations by replacing in the residual the correlation energy by individual pair energies, e.g., CEPA-2

$$R_{ab}^{ab} = V_{ab}^{ij} - \epsilon_{ij} T_{ab}^{ij} \; , \tag{109}$$

with

$$\epsilon_{ij} = (2 - \delta_{ij}) \sum_{ab} K_{ab}^{ij} \tilde{T}_{ab}^{ij} \; . \tag{110}$$

Other CEPA variants use slightly different expressions for the residual. The CEPA correlation energy is the sum of all pair energies

$$\mathcal{E}^{\rm CEPA} = \sum_{i \geq j} \epsilon_{ij} \; . \tag{111}$$

Obviously, the computational effort per iteration is virtually the same as for CISD, but the results are much better (almost as good as for CCSD(T) if singles are included).

2.5 Singles and doubles coupled-cluster

The main disadvantage of the variational configuration interaction method is the fact that it is not size consistent. This can easily be understood by considering two independent subsystems, e.g., two water molecules. The correct wavefunction for the total system AB should then be the (antisymmetrized) product of the wavefunctions of the two molecules A and B. If each of these wavefunctions contains double excitations from the SCF determinant, the total system will contain quadruple excitations, e.g.,

$$\Psi(A) = \Phi^{\rm SCF}(A) + \Psi^c(A) \;=\; [1 + \frac{1}{2} \sum_{ij}^{(A)} \sum_{ab}^{(A)} T_{ij}^{ab} \hat{E}_{ai} \hat{E}_{bj}] \Phi^{\rm SCF}(A)$$

$$\Psi(B) = \Phi^{\rm SCF}(B) + \Psi^c(B) \;=\; [1 + \frac{1}{2} \sum_{kl}^{(B)} \sum_{cd}^{(B)} T_{kl}^{cd} \hat{E}_{ck} \hat{E}_{dl}] \Phi^{\rm SCF}(B)$$

$$\begin{aligned} \Psi(AB) = \Phi^{\rm SCF}(AB) &+ \hat{\mathcal{A}}[\Phi^{\rm SCF}(A) \Psi^c(B) + \Phi^{\rm SCF}(B) \Psi^c(A)] \\ &+ \frac{1}{4} \sum_{ij}^{(A)} \sum_{ab}^{(A)} \sum_{kl}^{(B)} \sum_{cd}^{(B)} T_{ij}^{ab} T_{kl}^{cd} \hat{E}_{ai} \hat{E}_{bj} \hat{E}_{ck} \hat{E}_{dl} \Phi^{\rm SCF}(AB) \end{aligned} \tag{112}$$

where $\hat{\mathcal{A}}$ is the antisymmetrizer. It is seen that that the coefficients of the quadruple excitations $\Phi_{ijkl}^{abcd} = \hat{E}_{ai} \hat{E}_{bj} \hat{E}_{ck} \hat{E}_{dl} \Phi^{\rm SCF}(AB)$ are simple products $T_{ij}^{ab} T_{kl}^{cd}$ of the coefficients of the subsystems. However, these terms are not included in the CISD wavefunction for the dimer, and therefore the total CISD energy is not equal to the sum of the monomer energies.

In coupled-cluster theory[28,29,30] the wavefunction is generated by an exponential excitation operator

$$\Psi^{CC} = \exp(\hat{T})\Phi^{\rm SCF} \ , \tag{113}$$

where the exponential is defined by the Taylor expansion

$$\exp(\hat{T}) = 1 + \hat{T} + \frac{1}{2!}\hat{T}\hat{T} + \frac{1}{3!}\hat{T}\hat{T}\hat{T} + \dots \ . \tag{114}$$

The excitation operator $\hat{T}$ may be decomposed into single, double, and possibly higher excitation operators

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \dots \tag{115}$$

with

$$\hat{T}_1 = \sum_{ai} t_{ai}\hat{E}_{ai} \ , \tag{116}$$

$$\hat{T}_2 = \frac{1}{2}\sum_{ij}\sum_{ab} T^{ij}_{ab}\hat{E}_{ai}\hat{E}_{bj} \ , \tag{117}$$

$$\tag{118}$$

etc. Truncating the expansion after $\hat{T}_2$ yields the CCSD theory[31,21,32,22].

For two independent subsystems we can decompose $\hat{T}$ into a sum of two operators each acting only on one subsystem

$$\begin{aligned}\Psi(AB) = \exp(\hat{T}_A + \hat{T}_B)\Phi^{\rm SCF}(AB) \ &= \ \hat{\mathcal{A}}\left[\exp(\hat{T}_A)\Phi^{\rm SCF}(A)\exp(\hat{T}_B)\Phi^{\rm SCF}(B)\right] \\ &= \hat{\mathcal{A}}\left[\Psi(A)\Psi(B)\right] \ . \end{aligned} \tag{119}$$

Thus, the coupled-cluster wavefunction is size consistent as required. It implicitly contains triple, quadruple, and higher excitations, but the coefficients of these are all products of the single and double excitation amplitudes t^i_a and T^{ij}_{ab}.

Unfortunately, it is not possible to determine these amplitues variationally, since like the full CI expansion (113) includes up to N-fold excitations, which makes the evaluation of an expectation value too expensive. However, one can obtain a nonlinear system of equations for the amplitudes by projecting the Schrödinger equation from the left with the contravariant configurations $\tilde{\Phi}^a_i$ and $\tilde{\Phi}^{ab}_{ij}$ as defined in section 2.1. An additional equation for the correlation energy is obtained by projecting with the reference function. This yields

$$\mathcal{E}^{\rm CCSD} = \langle 0|\hat{H}(1 + \hat{T}_1 + \hat{T}_2 + \frac{1}{2}\hat{T}_1^2)|0\rangle \tag{120}$$

$$r^i_a = \langle\ \Phi^a_i|(\hat{H} - E^{\rm CCSD})(1 + \hat{T}_1 + \hat{T}_2 + \frac{1}{2}\hat{T}_1^2 + \hat{T}_1\hat{T}_2 + \frac{1}{3!}\hat{T}_1^3)|0\rangle \ = \ 0 \tag{121}$$

$$\begin{aligned} R^{ij}_{ab} = \langle\tilde{\Phi}^{ab}_{ij}|(\hat{H} - E^{\rm CCSD})(1 + \hat{T}_1 + \hat{T}_2 + \frac{1}{2}\hat{T}_1^2 + \hat{T}_1\hat{T}_2 + \frac{1}{3!}\hat{T}_1^3 \\ + \frac{1}{2}\hat{T}_1^2\hat{T}_2 \quad + \frac{1}{2}\hat{T}_2^2 + \frac{1}{4!}\hat{T}_1^4)|0\rangle \ = \ 0 \qquad (i \geq j, \ \text{all } a, b) \ . \end{aligned} \tag{122}$$

The expansions on the right-hand side terminate after the quadruple excitations since the hamiltonian can couple only configurations that differ by at most two excitations. The number of equations corresponds exactly to the number of amplitudes. Even though these equations look quite complicated, it turns out that their solution is not much more difficult than of the CISD equations. It can be shown that in the coupled-cluster case the contributions of the energy in the residual equations cancel out, as required for a size-consistent theory.

In order to exemplify the structure of the resulting equations, we will omit the single excitation operator $\hat{T}_1$ and consider only the coupled-cluster doubles (CCD) case. The full CCSD equations in a similar matrix formulation can be found in Ref. [22]. The explicit expressions for the CCD residual matrices $\mathbf{R}^{ij}$ are

$$\mathbf{R}^{ij} = \mathbf{K}^{ij} + \mathbf{K}(\mathbf{T}^{ij}) \; + \; \sum_{kl} \alpha_{ij,kl} \mathbf{T}^{kl} + \mathbf{G}^{ij} + \mathbf{G}^{ji} \; , \tag{123}$$

with

$$\mathbf{G}^{ij} = \mathbf{T}^{ij}\mathbf{X} - \sum_k \left[\beta_{ik} \mathbf{T}^{kj} - \tilde{\mathbf{T}}^{ik} \mathbf{Y}^{kj} + \frac{1}{2} \mathbf{T}^{ki} \mathbf{Z}^{kj} + (\mathbf{T}^{ki} \mathbf{Z}^{kj})^\dagger \right] \; . \tag{124}$$

The form of these equations is exactly the same as for the CID, discussed in the previous section, but there are now intermediate quantities which depend linearly on the amplitudes. In detail, the integrals K_{kl}^{ij} in the CID equations are replaced by $\alpha_{ij,kl}$, f_{ik} by β_{ik}, $\mathbf{f}$ by $\mathbf{X}$, $\mathbf{K}^{kj}$ by $\mathbf{Y}^{kj}$, and $\mathbf{J}^{kj}$ by $\mathbf{Z}^{kj}$. The explicit form of these quantites is

$$\alpha_{ij,kl} = K_{ij}^{kl} + \mathrm{tr}\left(\mathbf{T}^{ij}\mathbf{K}^{lk}\right) \; , \tag{125}$$

$$\beta_{ik} = f_{ik} + \sum_l \mathrm{tr}\left(\mathbf{T}^{il}\mathbf{L}^{lk}\right) \; , \tag{126}$$

$$\mathbf{X} \; = \mathbf{f} - \sum_{kl} \mathbf{L}^{kl}\mathbf{T}^{lk} \tag{127}$$

$$\mathbf{Y}^{kj} = \mathbf{K}^{kj} - \frac{1}{2}\mathbf{J}^{kj} + \frac{1}{4}\sum_l \mathbf{L}^{kl}\tilde{\mathbf{T}}^{lj} \; , \tag{128}$$

$$\mathbf{Z}^{kj} = \mathbf{J}^{kj} - \frac{1}{2}\sum_l \mathbf{K}^{lk}\mathbf{T}^{jl} \; . \tag{129}$$

The computational effort of the CCD differs from CID basically by the additional $2m^3$ matrix multiplications in eqs. (128) and (129), which doubles the time for evaluating the matrices $\mathbf{G}^{ij}$. However, the same external exchange operators $\mathbf{K}(\mathbf{T}^{ij})$ are needed, and therefore the difference in total time is less significant.

If singles are included, there are additional terms in the intermediates, but these require only minor computational effort. The products of singles arising from the $\hat{T}_1^2$, $\hat{T}_1^3$, and $\hat{T}_1^4$ terms in eqs. (121) and (122) can all be accounted for by defining modified amplitude matrices

$$\mathbf{C}^{ij} = \mathbf{T}^{ij} + \mathbf{t}^i \mathbf{t}^{j\,\dagger} \; , \qquad \bar{\mathbf{C}^{ij}} \; = \; \frac{1}{2}\mathbf{T}^{ij} + \mathbf{t}^i \mathbf{t}^{j\,\dagger} \; , \tag{130}$$

and then all intermediates depend only linearly on either $\mathbf{T}^{ij}$, $\mathbf{C}^{ij}$, or $\bar{\mathbf{C}}^{ij}$. The most notable difference between CISD and CCSD is that in the latter case one needs additional contractions of singles amplitudes with 3-external integrals

$$\mathbf{J}(\mathbf{E}^{ij})_{ab} = \sum_c (ab|ci)\, t_c^j \,, \tag{131}$$

$$\mathbf{K}(\mathbf{E}^{ij})_{ab} = \sum_c (ai|bc)\, t_c^j \,. \tag{132}$$

As the external exchange operators, these terms can be evaluated in two different ways. Either the 3-external integrals $(ab|ci)$ are explicitly generated, which requires a more expensive integral transformation (note, however, that the effort for the first quarter transformation is the same). Alternatively, the storage of these integrals can be avoided by computing these terms directly from the integrals in the AO basis. First, the singles amplitudes are transformed into the AO basis

$$t_\sigma^i = \sum_c X_{\sigma c}\, t_c^i \,, \tag{133}$$

then the operators are computed in the AO basis

$$\mathbf{J}(\mathbf{E}^{ij})_{\mu\nu} = \sum_\rho X_{\rho i} \sum_\sigma t_\sigma^j\, (\mu\nu|\rho\sigma) \,, \tag{134}$$

$$\mathbf{K}(\mathbf{E}^{ij})_{\mu\nu} = \sum_\rho X_{\rho i} \sum_\sigma t_\sigma^j\, (\mu\rho|\sigma\nu) \,, \tag{135}$$

and finally they are back transformed into the MO basis

$$\mathbf{J}(\mathbf{E}^{ij})_{\mathrm{MO}} = \mathbf{X}^\dagger \mathbf{J}(\mathbf{E}^{ij})_{\mathrm{AO}} \mathbf{X} \,, \tag{136}$$

$$\mathbf{K}(\mathbf{E}^{ij})_{\mathrm{MO}} = \mathbf{X}^\dagger \mathbf{K}(\mathbf{E}^{ij})_{\mathrm{AO}} \mathbf{X} \,. \tag{137}$$

This procedure, which is similar to the computation of the operators $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$, requires about $\frac{3}{4}m^4 m_{\mathrm{occ}} + 4m^3 m_{\mathrm{occ}}^2$ additions and multiplications (m basis functions, m_{occ} correlated orbitals) rather than $\frac{3}{2}m^3 m_{\mathrm{occ}}^2$ operations if the same quantities are computed from the fully transformed two-electron integrals (the full integral transformation scales as m^5). The additional effort is, however, quite insignificant as compared to the $\frac{1}{2}m^4 m_{\mathrm{occ}}^2$ operations needed to evaluate the operators $\mathbf{K}(\mathbf{T}^{ij})$ and will therefore not introduce a bottleneck. Nevertheless, it should be noted that the three-external integrals $(ab|ci)$ are also needed for evaluating the perturbative correction for triple excitations, and then it is of course advantageous to use them also for the CCSD.

Finally, we note that the QCISD (quadratic configuration interaction) equations[33] are obtained by omitting all $\hat{T}_1^2$, $\hat{T}_1^3$, $\hat{T}_1^4$ terms and the $\hat{T}_1\hat{T}_2$ term in equation 122. The residuals then include only part of the singles terms present in the CCSD. Most notably, the operators $\mathbf{J}(\mathbf{E}^{ij})$ and $\mathbf{K}(\mathbf{E}^{ij})$ are not needed in QCISD; as in the case of CISD all contributions of three-external integrals can be absorbed into the external exchange operators by computing these with modified coefficient matrices[22]. Another variant is the *Brueckner* coupled-cluster doubles

(BCCD) theory[34,35,36,37,38,39,22]. In this case the orbitals are modified in each iteration so that at convergence all singles amplitudes vanish. This can be achieved by aborbing after each update the singles into the orbitals

$$\phi_i \leftarrow \phi_i + \sum_a t_a^i \phi_a \tag{138}$$

with subsequent symmetrical reorthonormalization of the new occupied orbitals. Furthermore, the virtual orbitals have to be Schmidt-orthogonalized to the occupied space. Then the integral transformation must be repeated, since the $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$ change. The Brueckner theory has some theoretical advantages. In particular, the resulting wavefunction is less sensitive to symmetry breaking problems than the CCSD wavefunction on the basis of canonical Hartree-Fock orbitals.

2.6 Computational aspects

As already pointed out, the matrix formulation with a minimum amount of logic is one of the prerequisites for an efficient CISD or CCSD program. Often this can be exploited to the best possible extent by using highly optimzed routines for matrix multiplication (e.g, `dgemm`), which are available in BLAS (basic linear algebra subroutines) libraries on many platforms. These routines also allow to transpose one or both of the two matrices to be multiplied on the fly, without the need to precompute and store the transposed matrix. This is often useful, since the amplitudes $\mathbf{T}^{ij}$ are stored only for $i \geq j$, and $\mathbf{T}^{ji}$ is the transpose of $\mathbf{T}^{ij}$. The same holds for the operators $\mathbf{K}^{kl} = \mathbf{K}^{lk\dagger}$.

It is equally important to think carefully about memory and I/O usage. The number of amplitudes $\mathbf{T}^{ij}$, as well as the number of transformed integrals $\mathbf{J}^{kl}$, $\mathbf{K}^{kl}$ scale with the fourth power of the molecular size, and in large calculations it will often not be possible to keep all these quantities simultaneously in high speed memory. One can then use *paging algorithms*, which read blocks of data from disk as required. The algorithm should therefore be optimized so that for a given amount of available memory the I/O is minimized.

As a first example consider the evaluation of the matrices $\mathbf{G}^{ij}$ in the CID case. The $\mathbf{G}^{ij}$ do not need to be stored but their contribution can be immediately added to the residuals $\mathbf{R}^{ij}$. If the outer two loops run over j and k, one $\mathbf{J}^{kj}$ and one $\mathbf{K}^{kj}$ at a time need to be in memory and have to be read just once for a given kj. The simplest algorithm would then assume that all $\mathbf{R}^{ij}$ and $\mathbf{T}^{ik}$ can be kept in memory. Should this not be possible, one could split them into batches. For instance, if k is the outermost loop, one could read in this loop all $\mathbf{T}^{ik}$ for a fixed k; if still not all $\mathbf{R}^{ij}$ fit into memory, one could treat the largest possible subsets of them together. In this case, one would have to read the $\mathbf{J}^{kj}$, $\mathbf{K}^{kj}$, and $\mathbf{T}^{ik}$ for each batch of $\mathbf{R}^{ij}$. Reading all the $\mathbf{J}^{kj}$ and $\mathbf{K}^{kj}$ for each batch of $\mathbf{R}^{ij}$ could be avoided if each batch would comprise only a subset of j.

The situation is more complicated in the coupled cluster case, since then one has to evaluate the intermediates $\mathbf{Y}^{kj}$ and $\mathbf{Z}^{kj}$ instead of simply reading the $\mathbf{J}^{kj}$ and $\mathbf{K}^{kj}$. This requires all operators $\mathbf{K}^{kl}$ for a fixed k and all $\mathbf{T}^{lj}$ for fixed j. Thus, the simplest algorithm requires to keep all $\mathbf{R}^{ij}$ and $\mathbf{T}^{ij}$ together with all $\mathbf{K}^{kj}$ for a fixed k in memory. A simple paging over the $\mathbf{R}^{ij}$ and/or $\mathbf{T}^{ij}$ as in the CI case

is not possible, since this would involve repeated calculation of the intermediate quantities. It would be possible, however, first to evaluate the the $\mathbf{Y}^{kj}$ and $\mathbf{Z}^{kj}$, using a similar paging algorithm as in the CI case, and store these on disk. The $\mathbf{R}^{ij}$ are then computed in a second stage, exactly as in the CI case, but instead of the $\mathbf{J}^{kj}$ and $\mathbf{K}^{kj}$ one would read $\mathbf{Y}^{kj}$ and $\mathbf{Z}^{kj}$.

The computation time and memory requirements can be much reduced if molecular symmetry is exploited, which is easy as long as only one-dimensional irreducible representations are present, i.e. D_{2h} and subgroups. If symmetry adapted molecular orbitals are used, all matrices are blocked. The block structure of a given matrix T_{ab}^{ij} is determined by the product symmetry of the orbitals i and j, which must be the same as the product symmetry of a and b. The same holds for the $\mathbf{R}^{ij}$, $\mathbf{J}^{ij}$, and $\mathbf{K}^{ij}$. Of course, only the non-zero blocks are stored, and since each symmetry block can have a different dimension, the matrices are stored in one-dimensional arrays; block dimensions and offsets are precomputed and kept in memory. It is then convenient to have a set of subroutines for operations like matrix multiplications, matrix traces, outer products etc., which handle all the symmetry blocking internally. Thus, the rest of the program requires only a minimum amount of the symmetry information, and stays most readable and easy to debug.

2.7 Triple excitations

The accuracy of coupled cluster calculations with single and double excitations (CCSD) can be significantly improved by subsequently computing the effects of higher order excitations through Rayleigh-Schrödinger perturbation theory (RSPT) based on the Fock (Møller-Plesset) hamiltonian and the computed CCSD amplitudes of single and double excitations[40,33,41]. The most important such correction is that which is linear in triple excitations, since its inclusion gives an energy expression which is consistent with the exact solution of Schrödinger's equation up to fourth order[41,42,43,44]. The most widely used ansatz of this type, usually denoted CCSD(T)[41], is also consistent with many of the fifth order terms, and includes much of the sixth and higher order energies as well[45,46], provided that the reference wavefunction is a true variational solution of the Hartree-Fock equations. This analysis takes into account the fact that terms such as T_1T_2 present in the CCSD expansion already partially includes the effects of triple excitations.

The evaluation of the triples (T) correction requires terms like

$$W_{abc}^{ijk} = \sum_d (bd|ck)T_{ad}^{ij} - \sum_m (mj|ck)T_{ab}^{im} + \text{ permutations.} \tag{139}$$

The first terms scales with $m_{val}^3 m_{ext}^4$, the second with $m_{val}^4 m_{ext}^3$, where m_{val} and m_{ext} are the number of correlated and virtual (external) orbitals, respectively. Thus, the computational cost increases with $\mathcal{O}(\mathcal{N}^7)$, where $\mathcal{N}$ is a measure of the molecular size. In most cases the calculation of the triples correction is therefore much more expensive than the CCSD calculation itself, and the applicability is limited to quite small molecules. The elapsed time (not the cost!) can be reduced by parallelization of the code, but is should be noted that this does not substantially increase the molecular size that can be handled. Doubling the molecular size increases the time by a factor of 128, and therefore even the largest parallel computers

Table 1. CPU times[a] of coupled cluster calculations for glycine peptides[b]

Program	$(Gly)_1$	$(Gly)_2$	$(Gly)_3$
Basis functions	95	166	237
Transformation[c]	10	180	1471
CCSD (11 iterations)	312	7453	62741
Triples (T) correction	520	21081	220486

a) In seconds on Sun Enterprise 3500, Ultrasparc 336 MHZ processor
b) Using C_s symmetry
c) Partial transformation to generate two-external integrals $\mathbf{J}^{kl}$, $\mathbf{K}^{kl}$ and the three-external integrals $(ab|ci)$.

do not help much further. The dramatic increase of CPU time with molecular size is demonstrated in Table 1 for some glycine peptides, $(Gly)_n \equiv HO[C(O)CH_2NH]_nH$, using the correlation consistent double zeta basis set (cc-pVDZ) of Dunning[19]. The increase of the CPU times is close to the expected theoretical factors. It is easily estimated that the evaluation of the triples correction for the next larger peptide $(Gly)_4$ would already take about three weeks of CPU time. Another bottleneck of the triples calculation is the storage of the integrals $(ab|ci)$ over three external and one occupied orbitals, which must be stored on disk. Since these integrals have less permutational symmetry than the integrals in the AO basis, and the molecular orbitals are more diffuse than the basis functions, the number of significant integrals may even be larger than the number of AO integrals.

The cc-pVDZ basis set used in these calculations is too small for obtaining reliable results. Table 2 shows the dependence of the CPU times on the basis set for closed-shell coupled-cluster calculations on another molecule, p-dimethylbenzene C_8H_{10}, performed in C_s symmetry on a medium workstation. It is seen that increasing the basis set by about a factor of 1.6 increases the CPU times by a factor of 8-12, as expected from the quartic dependence. The larger calculation does not even include f-functions on the carbon atoms, as would be required for accurate results. The computation time is strongly dominated by the triples correction, while the differences of the various methods are quite small. Clearly, the treatment of molecules of this size is about the maximum what can be done in a reasonable time, which demonstrates the limitations of the conventional coupled cluster methods. Even the fastest current workstations or supercomputers are only about a factor of 3-4 faster, and do not much extend the range of applicability. The strong dependence of the computer time on the molecular size can be dramatically reduced using local correlation methods, as will be discussed in section 4. In particular, as will be demonstrated in section 4.3, the evaluation of an approximate local triples corrections no longer dominates the calculation, but takes only a small amount of the total time.

Table 2. CPU times[a] of coupled cluster calculations for C_8H_{10} with different basis sets

Program	cc-pVDZ[b]	cc-pVTZ(d/p)[c]
Transformation[d]	35	318
CCSD/iteration	374	2313
QCISD/iteration	360	2180
BCCD/iteration	399	2520
Transformation[e]	119	1443
Triples (T) correction	9059	122515

a) In seconds on HP J282, PA8000/180MHZ processor
b) 162 basis functions ($114a', 48a''$)
c) 274 basis functions ($188a', 86a''$)
d) Partial transformation to generate the two-external integrals $\mathbf{J}^{kl}$, $\mathbf{K}^{kl}$
e) Partial transformation to generate two-external integrals $\mathbf{J}^{kl}$, $\mathbf{K}^{kl}$ and the three-external integrals $(ab|ci)$

3 Open-shell single-reference methods

The coupled-cluster treatment of open-shell systems is more complicated that the closed shell case since additional types of orbitals and excitations occur. First of all, it is possible to use either a spin-unrestricted (UHF) or a spin-restricted (RHF) Hartree-Fock wavefunction as a reference. In the UHF case the α and β spin orbitals are optimized independently, which leads to a wavefunction that is not an eigenfunction of the total spin operator $\hat{S}^2$. It is well known that the problems associated with the spin-contamination of the UHF wavefunction can become magnified when electron correlation effects are introduced[1], in particular in second-order perturbation theory (UMP2). It is therefore more desirable to use RHF orbitals.

The second difficulty is the definition of the excitation operators used in coupled-cluster treatments. It turns out that a fully spin-adapted treatment based on an RHF reference function and the spin-free excitation operators $\hat{E}_{rs}$ is very complicated. Is is much easier to use spin-orbital excitation operators $\hat{e}_{ai}$, which replace a spin-orbital ψ_i by another spin orbital ψ_a with the same spin. However, then the correlated wavefunction is not spin-adapted, even if an RHF reference functions is used. This problem already arises in the linear configuration interaction theory if the first order interacting space, spanned by the functions $\hat{e}_{ai}\hat{e}_{bj}|\Psi_0\rangle$, is used as a basis; this is due to the fact that for high-spin open shell cases this space does not include all possible Slater determinants of given M_S which arise from a particular occupancy of spatial orbitals. For instance, in a three electron case with reference function $|\phi_1^\alpha\phi_1^\beta\phi_2^\alpha|$, the determinant $|\phi_a^\alpha\phi_b^\alpha\phi_2^\beta|$ is a triple excitation and not included in the first order interacting space. This function would be necessary, however, to generate one of the two possible doublet spin eigenfunctions together with the determinants $|\phi_a^\beta\phi_b^\alpha\phi_2^\alpha|$ and $|\phi_a^\alpha\phi_b^\beta\phi_2^\alpha|$. A quartet spin contamination arises if the

latter two Slater determinants have coefficients of different magnitude. Thus, the RHF-UCISD and RHF-UCCSD theories based on spin-orbital single and double excitations are not spin adapted.

As will be shown in Section 3.2, the spin contamination in the linear UCISD wavefunction can be quite easily removed by applying appropriate projection operators to the UCISD residual vector. The same projection can be used to remove the spin contamination from the linear terms of the CCSD wavefunction. But even then, the presence of higher powers of $\hat{T}$ in the CCSD can introduce a spin contamination in a non–trivial way. Fortunately, this effect is usually very small. The partial spin adpation (PSA-CCSD) of only the linear terms has a number of advantages: the number of independent parameters (amplitudes) is minimized and corresponds exactly to the first-order interacting space; also spin contamination effects are minimized, though not entirely removed. In an optimum implementation, the computational cost of the PSA-CCSD should be approximately the same as for a closed shell calculation with the same number of correlated orbitals.

3.1 Spin-unrestricted coupled-cluster theory (UCCSD)

We will first consider the spin unrestricted coupled cluster (UCCSD) for the case that the reference function is a high-spin RHF Slater determinant with m_{closed} doubly occupied and m_{open} singly occupied orbitals; high spin means that all open-shell electrons have α spin. The UCCSD wavefunction is obtained using the following cluster operator $\hat{T} = \hat{T}_1 + \hat{T}_2$ in the exponential ansatz (113)

$$\begin{aligned}\hat{T} = &\sum_{ia} (\tilde{t}^i_a \hat{e}^\alpha_{ai} + \bar{t}^i_a \hat{e}^\beta_{ai}) + \sum_{it} \bar{t}^i_t \hat{e}^\beta_{ti} + \sum_{ta} \tilde{t}^t_a \hat{e}^\alpha_{at} + \sum_{ij}\sum_{ab} (\tilde{T}^{ij}_{ab} \hat{e}^\alpha_{ai} \hat{e}^\alpha_{bj} + \bar{T}^{ij}_{ab} \hat{e}^\beta_{ai} \hat{e}^\beta_{bj}) \\ &+ \sum_{ij}\sum_{ab} T^{ij}_{ab} \hat{e}^\alpha_{ai} \hat{e}^\beta_{bj} + \sum_{ij}\sum_{at} T^{ij}_{at} \hat{e}^\alpha_{ai} \hat{e}^\beta_{tj} + \sum_{ij}\sum_{tu} \bar{T}^{tu}_{ab} \hat{e}^\beta_{ti} \hat{e}^\beta_{uj} \\ &+ \sum_{tj}\sum_{ab} T^{tj}_{ab} \hat{e}^\alpha_{at} \hat{e}^\beta_{bj} + \sum_{tj}\sum_{au} T^{tj}_{au} \hat{e}^\alpha_{at} \hat{e}^\beta_{uj} + \sum_{tu}\sum_{ab} \tilde{T}^{tu}_{ab} \hat{e}^\alpha_{at} \hat{e}^\alpha_{bu} \,, \end{aligned} \tag{140}$$

where $\hat{e}^\sigma_{ai} = \hat{\eta}^{\sigma\dagger}_a \hat{\eta}^\sigma_i$ are the usual spin-orbital excitation operators. If applied to a Slater determinant, $\hat{e}^\sigma_{ai}$ replaces spin orbital ϕ^σ_i by ϕ^σ_a; $\sigma = \{\alpha, \beta\}$ denotes the spin. Here and in the following, the indices i, j refer to closed-shell orbitals, t, u to open-shell orbitals, and a, b to virtual orbitals. For each orbital pair (ij), there are three sets of amplitudes, namely those for pure α or β-spin excitations $\tilde{T}^{ij}_{ab}$ and $\bar{T}^{ij}_{ab}$, respectively, and those for mixed α, β excitations T^{ij}_{ab}. In total, there are about three times as many amplitudes as in the closed-shell case. The corresponding cluster amplitudes are obtained by solving a non-linear set of equations obtained by projecting the Schrödinger equation on the left with $\Psi^{\rm RHF} \equiv |0\rangle$, $\hat{e}^\sigma_{ai}|0\rangle$, $\hat{e}^\sigma_{ai}\hat{e}^{\sigma'}_{bj}|0\rangle$ etc., as in eqs. (120) - (122). The resulting explicit equations can be found in Ref. [47]. They have a very similar matrix structure as the closed shell equations discussed in the previous sections and will not be further discussed here. It should be noted, however, that there are three times as many equations as in the closed shell case, and the total computational effort is about three times larger.

3.2 Partially spin-resticted coupled-cluster theory (RCCSD)

In fully spin coupled theory[48], it is recognized that the hamiltonian operator is spin free, and therefore the excitation operators used in the previous section may be replaced by the smaller set $\hat{E}_{ai}, \hat{E}_{at}, \hat{E}_{ti}$ and their products, where again $t, u, \ldots$ are used to denote orbitals lying in the singly occupied space, while $i, j, \ldots$ denote true closed shell orbitals, and $a, b, \ldots$ external orbitals. A simpler theory[47,49,50,51], including some but not all of the spin coupling, may be obtained by using the operators $\hat{E}_{ai}, \hat{e}^{\alpha}_{at}, \hat{e}^{\beta}_{ti}$ and their products; because the orbitals ϕ^{α}_{t} are occupied, and ϕ^{β}_{t} are unoccupied in Ψ_0, the wave function is then be spin adapted for a CISD configuration expansion, which is linear in these operators. In the non-linear CCSD case products of these operators can still give a spin-contaminated contribution to the wave function. This ansatz is denoted "partially spin adapted" CCSD (PSA-CCSD). It has the advantage that the complications occuring through the spin adaption are minimized, while most of the spin-contamination is removed.

A slight complication arises for the so called *semi-internal* configurations generated by the operators $\hat{e}^{\alpha}_{at}\hat{e}^{\beta}_{ti}$, which have the same orbital occupancy as the single excitations $\hat{E}_{ai}$ but a different spin contribution. It is easily seen that $\hat{e}^{\alpha}_{at}\hat{e}^{\beta}_{ti}|0\rangle$ is not a spin eigenfunction; a correct spin eigenfunction is generated by the operator $\hat{e}^{\alpha}_{at}\hat{e}^{\beta}_{ti} - \frac{1}{2}\hat{e}^{\alpha}_{ai} + \frac{1}{2}\hat{e}^{\beta}_{ai}$. In fact, analysis of the action of the hamiltonian operator on the RHF reference function shows that this operator together with $\hat{E}_{ai}$ generates the two possible spin eigenfunctions that contribute to the first-order interacting space. The cluster operator can now be written as

$$\begin{aligned}\hat{T} = &\sum_{ia} (\tilde{t}^{i}_{a}\hat{e}^{\alpha}_{ai} + \bar{t}^{i}_{a}\hat{e}^{\beta}_{ai}) + \sum_{it} \bar{t}^{i}_{t}\hat{e}^{\beta}_{ti} + \sum_{ta} \tilde{t}^{t}_{a}\hat{e}^{\alpha}_{at} \\ &+ \sum_{ij}\sum_{ab} T^{ij}_{ab}\hat{E}_{ai}\hat{E}_{bj} + \sum_{ij}\sum_{at} T^{ij}_{at}\hat{E}_{ai}\hat{e}^{\beta}_{tj} + \sum_{tj}\sum_{ab} T^{tj}_{ab}\hat{e}^{\alpha}_{at}\hat{E}_{bj} \\ &+ \sum_{tj}\sum_{au} T^{tj}_{au}\hat{e}^{\alpha}_{at}\hat{e}^{\beta}_{uj} + \sum_{tu}\sum_{ab} T^{tu}_{ab}\hat{e}^{\alpha}_{at}\hat{e}^{\alpha}_{bu} \,, \end{aligned} \tag{141}$$

with the restrictions

$$\tilde{t}^{i}_{a} = t^{i}_{a} - \frac{1}{2}\sum_{t} T^{ti}_{at} \,, \tag{142}$$

$$\bar{t}^{i}_{a} = t^{i}_{a} + \frac{1}{2}\sum_{t} T^{ti}_{at} \,, \tag{143}$$

which account for the fact that there are only two independent spin eigenfunctions for the orbital configurations $\ldots\phi_i\phi_t\phi_a$, as discussed above. Equating the operator $\hat{T}$ with the spin-unrestricted operator in eq. (140) yields the following relations between the amplitudes

$$\tilde{T}^{tu}_{ab} = T^{tu}_{ab} \,, \tag{144}$$

$$\bar{T}^{ij}_{tu} = T^{ij}_{tu} \,, \tag{145}$$

$$\tilde{T}^{pj}_{ab} = T^{pj}_{ab} - T^{pj}_{ba} \,, \tag{146}$$

$$\bar{T}^{ij}_{ar} = T^{ij}_{ar} - T^{ji}_{ar} \,, \tag{147}$$

where p, q refer to all occupied orbitals (closed + open), and rs to all openshell + virtual orbitals. Setting further $t_t^i = \frac{1}{2}\bar{t}_t^i$ and $t_a^t = \frac{1}{2}\tilde{t}_a^t$ we obtain a unique set of amplitudes T_{rs}^{pq} and t_r^p to be solved for. The number of independent parameters is then exactly the same as in a fully spin adapted formulation and about three times smaller than in the spin-unrestricted case. The corresponding minimal set of coupled equations can be obtained by projecting the Schrödinger equation onto the set of functions generated the individual excitation operators in the cluster operator to the reference function. Since the configuration generated in this way are non-orthonormal, simpler equations can again be derived by projecting the Schrödinger equation with the equivalent set of contravariant configurations. For details refer to Ref. [47].

The simplest possibility to solve the PSA-CCSD equations is to compute the UCCSD residuals, and then to form appropriate linear combinations of the different spin components to generate the spin-restricted residuals as needed for updating the amplitudes. Finally, the UCCSD amplitudes can be generated from the PSA-CCSD ones using eqs. (144–147). Of course, this procedure does not save any computer time relative to the UCCSD, but it requires only a minor modification of an existing UCCSD program to perform the spin projection.

4 Linear scaling local correlation methods

As pointed out in the previous sections, the computational cost of conventional electron correlation methods like MP2 or CCSD(T) increases dramatically with the size of the system. The steep scaling mainly originates from the delocalized character of the canonical MO basis. This leads to a quadratic increase of the number of amplitudes used for correlating a given electron pair, and a quartic increase of the total number of parameters. The increase of the CPU time with molecular is even steeper, being $\mathcal{O}(\mathcal{N}^7)$ for the best method of choice, which is usually CCSD(T).

From a physical point of view, however, there should be no need to correlate all electrons in an extended molecular system: dynamic electron correlation in non-metallic systems is a short-range effect with an asymptotic distance dependence of $\propto r^{-6}$ (dispersion energy), and thus the high-order dependence of the computational cost with the number of electrons of the system is just an artifact of the canonical orthogonal basis, in which the diverse correlation methods have traditionally been formulated. One natural way to circumvent this problem is to to use *local* orbitals to span the occupied and virtual spaces. Such *local correlation methods* have been proposed by several authors. Some recent papers which also summarize previous work can be found Refs. [52,53,54,55,56].

Particularly successful has been the local correlation method originally proposed by Pulay[57], which was first implemented by Saebø and Pulay for Møller-Plesset perturbation theory up to fourth order (LMP2 - LMP4(SDQ) without triple excitations) and the coupled-electron pair approximation (CEPA)[58,59]. Later it was generalized to full local CCSD by Hampel and Werner [53]. While in the early work of Saebø and Pulay[58,59] it could already be shown that only 1-2% of the correlation energy (relative to a conventional calculation with the same basis set) is lost

by the local approximation, it was not yet possible at that time to demonstrate that the scaling of the computational cost can actually be reduced, and that larger systems than with conventional methods can be treated. Significant progress in this direction was only made during the last few years when the local correlation methods were combined with newly developed integral-direct techniques[60], which fully expoit the possibilities for integral screening. Within such a framework, it has been possible to develop $\mathcal{O}(\mathcal{N})$ algorithms (asymptotic linear scaling of all computational ressources, i.e. CPU time, memory and disk space with molecular size) for local MP2 [54], local CCSD [61] and even for local connected triples correction (T) [62].

In the local correlation methods the occupied space is usually spanned by *localized molecular orbitals* (LMOs), which are obtained from the occupied canonical orbitals of a preceeding SCF calculation by virtue of a unitary localization procedure [63,64,65], which maintains the orthogonality the occupied SCF orbitals[a]

$$|\phi_k^{\rm LOC}\rangle = \sum_i |\phi_i^{\rm CAN}\rangle W_{ik} \qquad \text{with} \qquad \mathbf{W}\mathbf{W}^\dagger = \mathbf{1} \; . \tag{148}$$

The corresponding MO coefficient matrices are related similarly

$$\mathbf{L} = \mathbf{X}_{\rm OCC}\mathbf{W} \; . \tag{149}$$

(If core orbitals are not correlated, the localization should be restriced to the subspace of correlated valence orbitals.) The idea of Pulay was to abandon the orthogonality of the virtual orbitals, and to use a basis of functions which resemble the atomic orbitals (AOs) as much as possible. Obviously, the AOs are optimally localized, but since they are not orthogonal on the occupied orbitals one cannot use them straightaway. The strong orthogonality between the occupied and virtual spaces must be retained, since otherwise excitations would violate the Pauli exclusion principle and the theory would become very complicated. The orthogonality to the occupied space can be enforced by applying a projection operator $(1 - \sum_i |\phi_i\rangle\langle\phi_i|)$ to the AOs, yielding projected atomic orbitals (PAOs)

$$\begin{aligned} |\tilde{\chi}_r\rangle &= (1 - \sum_{i=1}^{m_{occ}} |\phi_i\rangle\langle\phi_i|)|\chi_r\rangle \\ &= \sum_\mu |\chi_\mu\rangle P_{\mu r} \end{aligned} \tag{150}$$

with

$$\mathbf{P} = \mathbf{1} - \mathbf{L}\mathbf{L}^\dagger\mathbf{S} \;=\; \mathbf{1} - \mathbf{X}_{\rm OCC}\mathbf{X}_{\rm OCC}^\dagger\mathbf{S} \;=\; \mathbf{X}_{\rm VIRT}\mathbf{X}_{\rm VIRT}^\dagger\mathbf{S}. \tag{151}$$

Here, $\mathbf{X}_{\rm OCC}$ and $\mathbf{X}_{\rm VIRT}$ denote the rectangular submatrices of the MO coefficient matrix $\mathbf{X}$ for the occupied and virtual (external) canonical orbitals, respectively, i.e.,

$$(\mathbf{X}_{\rm VIRT}\mathbf{X}_{\rm VIRT}^\dagger)_{\mu\nu} = \sum_a X_{\mu a} X_{\nu a} \; , \tag{152}$$

[a] Recently, it has also been proposed to use non-orthogonal basis functions to span the occupied space[66,67], but the computational efficiency of this approach has not yet been proven.

and the last equality in Eq. (151) follows from the orthonormality condition

$$(\mathbf{X}_{\mathrm{OCC}}\mathbf{X}_{\mathrm{OCC}}^{\dagger} + \mathbf{X}_{\mathrm{VIRT}}\mathbf{X}_{\mathrm{VIRT}}^{\dagger})\mathbf{S} = \mathbf{1} \; . \tag{153}$$

The PAOs are orthogonal to all occupied orbitals

$$\langle \tilde{\chi}_r | \phi_i^{\mathrm{LOC}} \rangle = (\mathbf{P}^{\dagger}\mathbf{S}\mathbf{L})_{ri} \; = \; 0 \tag{154}$$

but non-orthogonal among themselves

$$\langle \tilde{\chi}_r | \tilde{\chi}_s \rangle = (\mathbf{P}^{\dagger}\mathbf{S}\mathbf{P})_{rs} \; = \; (\mathbf{S}\mathbf{X}_{\mathrm{VIRT}}\mathbf{X}_{\mathrm{VIRT}}^{\dagger}\mathbf{S})_{rs} \; . \tag{155}$$

For non-metallic systems the PAOs are intrinsically localized, though less well than the unprojected AOs. Due to the projection the full set of PAOs is linearly dependent, but these linear dependencies can be removed at a later stage.

After having introduced local functions to span both the occupied and the virtual spaces, it is possible to *truncate* the expansion of the wavefunction in a physically reasonable way. First, one assigns to each localized orbital ϕ_i^{LOC} an orbital domain $[i]$ which contains all AOs needed to approximate the orbital ϕ_i^{LOC} with a prescribed accuracy. In practice, always all AOs at a given atom are treated together, and as many atoms are added as required. The order in which atoms are added is determined by gross atomic Mulliken charges. The corresponding orbital domain in the virtual space is spanned by the PAOs generated by applying the projector to the selected AOs. The PAOs in domain $[i]$ are then all spatially close to the localized orbital ϕ_i^{LOC}. This selection procedure can be performed fully automatically as described in Ref. [68].

The first approximation to the correlated wavefunction is now that single excitations from orbital ϕ_i^{LOC} are resticted to PAOs in the domain $[i]$, while double excitations from a pair of occupied LMOs i and j are restricted to a subset $[ij]$ of PAOs. The *pair domain* $[ij]$ is simply the *union* of the two orbital domains $[i]$ and $[j]$. The immediate consequence of these truncations is that for a given pair ij the number of amplitudes T_{rs}^{ij}, $rs \in [ij]$ no longer increases quadratically with increasing molecular size, but instead becomes *independent* of molecular size.

The second approximation is to introduce a *hierarchical treatment* of different pairs based on the interorbital distance R_{ij} between two LMOs i and j. R_{ij} is defined as the shortest distance between any centre included in the orbital domain $[i]$ and any centre in the domain $[j]$. We distinguish *strong*, *weak*, *distant*, and *very distant* pairs. The strong pairs have at least one atom in common and usually account for about 95% of the correlation energy. These pairs are treated at highest level, e.g., CCSD. Weak pairs are those for which the minimum distance is smaller than typcially 8 bohr. These pairs can be treated at lower level, e.g., MP2. Distant pairs ($8 \leq R_{ij} \leq 15$ bohr) are also treated by MP2, but the required two-electron integrals can be approximated by a multipole expansion[69], which reduces the cost for the integral transformation (see section 6.3). Finally, the very distant pairs ($R_{ij} > 15$ bohr) contribute to the correlation energy only by a few micro hartree and can therefore be neglected. The important point to notice is now that the number of strong, weak, and distant pairs all scale linearly with size. Only the number of very distant pairs, which are neglected, scales quadratically. This is demonstrated in Fig. 4 for linear chains of glycine peptides, $(\mathrm{Gly})_n \equiv \mathrm{HO[C(O)CH_2NH]}_n\mathrm{H}$, Thus, the

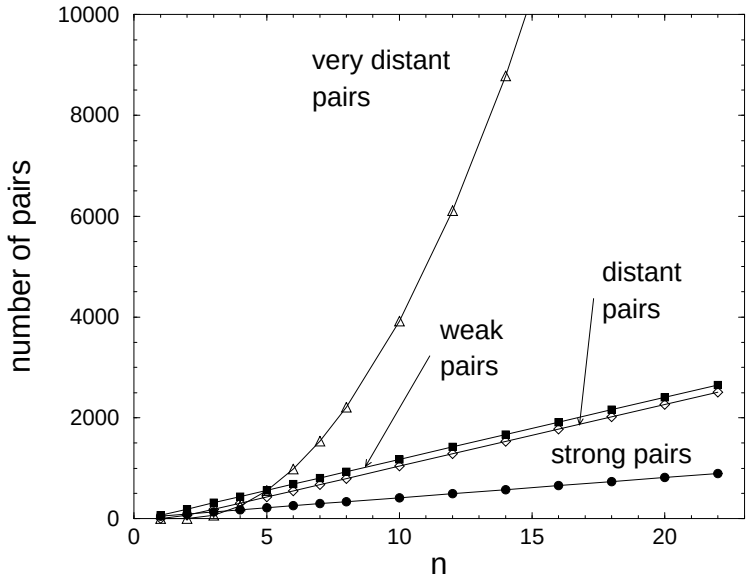


Figure 4. Number of pairs for a chain of Glycine peptides $(\mathrm{Gly})_n$ as function of the chain length.

total number of amplitudes on which the wavefunction depends scales only linearly with molecular size. This forms the basis for the development of electron correlation methods with linear cost scaling. Furthermore, the number of strong pairs remains quite modest, which is very important for an efficient CCSD algorithm (cf. section 4.2).

4.1 Local MP2

In the local LMO/PAO basis, the first-order wave function takes the form

$$|\Psi^{(1)}\rangle = \frac{1}{2} \sum_{ij \in P} \sum_{rs \in [ij]} T_{rs}^{ij} |\Phi_{ij}^{rs}\rangle \qquad \text{with } T_{rs}^{ij} = T_{sr}^{ji} \,, \tag{156}$$

where P represents the truncated pair list and it is implicitly assumed that the pair domains $[ij]$ are defined as described above. The configurations $|\Phi_{ij}^{rs}\rangle$ are defined as in eq. (49), but now the virtual labels r, s refer to the non-orthogonal PAOs. Note that the commutation relations of the excitation operators involving non-orthogonal orbitals are different and depend on overlap matrix elements.

In order to derive the LMP2 equations in the non-orthogonal basis of PAOs we first consider the transformation properties of the operators and amplitudes. The projected orbitals can be expressed in the basis of virtual orbitals as

$$\mathbf{P} = \mathbf{X}_{\mathrm{VIRT}} [\mathbf{X}_{\mathrm{VIRT}}^{\dagger} \mathbf{S}] \;=\; \mathbf{X}_{\mathrm{VIRT}} \mathbf{V} \,, \tag{157}$$

and therefore the MP2 residual given in eq. (79) for a basis of orthogonal MOs can be transformed to the PAO basis as

$$\mathbf{R}^{ij}_{\mathrm{PAO}} = \mathbf{V}^\dagger \mathbf{R}^{ij}_{\mathrm{MO}} \mathbf{V} \; . \tag{158}$$

The Fock and exchange matrices transform similarly. The transformation properties of the amplitude matrices can be obtained by expanding the projected orbitals in the pair correlation functions Ψ_{ij} into the MO basis

$$\Psi_{ij} = \sum_{rs\in[ij]} T^{ij}_{rs} |\ldots\tilde{\chi}_r\tilde{\chi}_s\ldots| \; = \; \sum_{ab} \left(\sum_{rs\in[ij]} V_{ar} T^{ij}_{rs} V_{bs} \right) |\ldots\phi_a\phi_b\ldots| \tag{159}$$

which yields the relation

$$\mathbf{T}^{ij}_{\mathrm{MO}} = \mathbf{V}\mathbf{T}^{ij}_{\mathrm{PAO}}\mathbf{V}^\dagger \; . \tag{160}$$

Inserting this into eq. (79) yields

$$\begin{aligned} \mathbf{R}^{ij}_{\mathrm{PAO}} = \mathbf{K}^{ij}_{\mathrm{PAO}} &+ \mathbf{f}_{\mathrm{PAO}}\mathbf{T}^{ij}_{\mathrm{PAO}}\mathbf{S}_{\mathrm{PAO}} + \mathbf{S}_{\mathrm{PAO}}\mathbf{T}^{ij}_{\mathrm{PAO}}\mathbf{f}_{\mathrm{PAO}} \\ &- \sum_k \mathbf{S}_{\mathrm{PAO}} \left[f_{ik}\mathbf{T}^{kj}_{\mathrm{PAO}} + f_{kj}\mathbf{T}^{ik}_{\mathrm{PAO}} \right] \mathbf{S}_{\mathrm{PAO}} = 0 \; , \end{aligned} \tag{161}$$

where $\mathbf{S}_{\mathrm{PAO}} = \mathbf{P}^\dagger\mathbf{S}_{\mathrm{AO}}\mathbf{P} = \mathbf{V}^\dagger\mathbf{V}$ is the overlap matrix of the projected orbitals, cf. eq. (155). In the local basis the occupied-occupied and virtual-virtual blocks of the Fock matrix are not diagonal, and therefore the linear equations (161) have to be solved iteratively for the amplitudes $\mathbf{T}^{ij}_{\mathrm{PAO}}$. Restricting the excitations to domains $[ij]$ of PAOs means that only the elements T^{ij}_{rs} with $r, s \in [ij]$ are nonzero, and only the corresponding elements of the residual, $R^{ij}_{rs},\ r, s \in [ij]$ must vanish at convergence. For a given set of amplitudes, the Hylleraas functional (eq. 92)

$$E_2 = \sum_{ij\in P} \sum_{rs\in[ij]} (2T^{ij}_{rs} - T^{ij}_{sr})(K^{ij}_{rs} + R^{ij}_{rs}) \tag{162}$$

can be computed. At convergence, $R^{ij}_{rs} = 0$ for $r, s \in [ij]$, and then $E_2 = E^{(2)}$.

Since the projected orbitals are not orthogonal and may even be linearly dependent, straightforward application of an update formular as eq. (103) will lead to slow or no convergence. In order to perform the amplitude update it is therefore necessary to transform the residuals to a pseudo-canonical basis, which diagonalizes the Fock operator in the subspace of the domain $[ij]$, i.e.

$$f^{ij}_{rs}X^{ij}_{rs} = S^{ij}_{rs}X^{ij}_{sa}\epsilon^{ij}_a \qquad \text{for} \qquad r, s \in [ij] \; , \tag{163}$$

$$R^{ij}_{ab} = \sum_{rs\in[ij]} X^{ij}_{sa} R^{ij}_{rs} X^{ij}_{sb} \; . \tag{164}$$

The update is then computed in this orthogonal basis and finally backtransformed to the projected basis

$$\Delta T^{ij}_{ab} = -R^{ij}_{ab}/(\epsilon^{ij}_a + \epsilon^{ij}_b - f_{ii} - f_{jj}) \; , \tag{165}$$

$$\Delta T^{ij}_{rs} = \sum_{ab} X^{ij}_{ra}\Delta T^{ij}_{ab} X^{ij}_{sb} \; . \tag{166}$$

Note that the square transformation matrix $\mathbf{X}^{ij}$ is different for each electron pair. The dimension of this matrix corresponds to the number of projected orbitals in domain $[ij]$ and is therefore independent of the molecular size. If the overlap matrix $S^{ij}_{rs},\ r,s \in [ij]$ has small or zero eigenvalues, i.e, if the functions in the domain are linearly dependent, the corresponding eigenvectors of $\mathbf{S}^{ij}$ are projected out[53]. Convergence of this scheme is reached quickly; usually 5-7 iterations are sufficient to converge the energy to better than 0.1 μH using no further convergence acceleration[70].

In order to compute the residuals, only the small subset of exchange integrals

$$K^{ij}_{rs} = (ri|sj) = \sum_{\nu\mu} P_{\mu r} P_{\nu s} \left[\sum_{\rho\sigma} L_{\rho i} L_{\sigma j} (\mu\rho|\nu\sigma) \right] \qquad r,s \in [ij] \qquad (167)$$

is needed, where all r,s are close either to i or j. This makes it possible to devise an integral-direct transformation scheme which scales only linearly with molecular size[54]. Taking further into account that for a given pair (ij) the number of terms k in the summation of eq. (161) becomes asymptotically independent of the molecular size (provided very distant pairs are neglected), it follows that the computational effort to solve the linear equations scales inearly with molecular size as well[54]. Thus, the overall cost to transform the integrals, to solve the linear equations (161), and to compute the second order energy depends linearly on the molecular size. This has made it possible to perform LMP2 calculations with about 2000 basis functions and 500 correlated electrons without using molecular symmetry. Since also the memory demands are small and scale linearly with molecular size, such calculations can even be performed on low-cost personal computers.

Finally we note that analytical energy gradients for LMP2 have been developed[71]. It has been shown that the local ansatz largely eliminates basis set superposition errors (BSSE), and it is therefore possible to optimize BSSE-free equilibrium structures of molecular clusters[72,73]. Recently, also the theory for computing NMR chemical shifts using the LMP2 method has been derived and first promising results have been obtained[74].

4.2 Local CCSD

The LCCSD equations can be obtained exactly in the same way as indicated above for the LMP2 case, namely by transforming the residuals from the MO to the PAO basis. The resulting equations differ formally from the canonical ones only by the occurrence of additional matrix multiplications with the overlap matrix. The full formalism has been presented in Ref. [53] and will therefore not be repeated here.

As already pointed out before, it is usually sufficient to treat pairs with interorbital distances $R_{ij} \leq 1$ bohr (strong pairs) at the CCSD level. Exceptions are cases where it is of importance to treat longe-range interactions accurately at high level, for instance for computing intermolecular interactions. In the following discussion we will assume, however, that this is not the case, and that the number of strong pairs included in the CCSD treatment is relatively small and scales linearly with molecular size, as shown in Fig. 4.

For the LMP2 case it is immediately obvious that the number of transformed exchange integrals $K_{rs}^{ij} = (ri|js)$ that need to be computed and stored depends only linearly on the molecular size. This follows from the fact that there is a one-to-one correspondence between these integrals and the corresponding amplitudes T_{rs}^{ij}. In the coupled cluster case however, the situation is more complicated, since integrals like the above also couple different electron pairs in the CCSD formalism. Furthermore, as already discussed in section 6.6, there are additional contributions of Coulomb integrals $J_{rs}^{ij} = (ij|rs)$, as well as of integrals $(ir|st)$ and $(rs|tu)$ with three and four external indices, respectively. Closer inspection of the problem reveals, however, that also in the coupled cluster case the number of transformed integrals scales only linearly with molecular size. The same is true for the number of floating point operations needed to compute the residuals.

In order to illustrate the main ideas we will a consider the contribution of the $\mathbf{Y}^{jk}$ intermediates to the LCCD residual, cf. eqs. (124) and (128),

$$\mathbf{G}^{ij} = \ldots + \sum_k \mathbf{S}\tilde{\mathbf{T}}^{ik} \left(\mathbf{K}^{kj} - \frac{1}{2}\mathbf{J}^{kj} \right) + \frac{1}{4} \sum_{kl} \mathbf{S}\tilde{\mathbf{T}}^{ik}\mathbf{L}^{kl}\tilde{\mathbf{T}}^{lj}\mathbf{S} + \ldots \,. \tag{168}$$

Here, all matrices are assumed to be in the PAO basis. Now, since (ik) and (lj) both are strong pairs, there is only a constant number of LMOs k and l interacting with given i and j, respectively. Furthermore, since also (ij) is a strong pair, it follows that for a fixed (ij) the total number of operators contributing to each $\mathbf{G}^{ij}$ is asymptotically constant and independent of the molecular size. Thus, the total number of integral matrices $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$ needed in eq. 168 scales linearly; the same holds for the number of matrix multiplications. Furthermore, the LMOs k and l of the surviving operators have to be close, which is important to achieve linear scaling in the integral transformation needed to compute the $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$. Note that fewer $\mathbf{J}^{kl}$ than $\mathbf{K}^{kl}$ are needed, since the $\mathbf{J}^{kl}$ only occur in the linear terms. Thus, separate operator lists for the $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$ have to be maintained. In contrast to the canonical case, the evaluation of the residuals is driven by individual $\mathbf{J}^{kl}$ and $\mathbf{K}^{kl}$, and the $\mathbf{Y}^{jk}$ and $\mathbf{Z}^{jk}$ intermediates are never explicitly computed.

The PAO range r, s of a particular operator K_{rs}^{kl} is also independent of molecular size: since i must be close to k, and l close to j, all the r, s occuring in the matrix multiplications of eq. 168 must be within a limited distance to k, l. This leads to a different *operator domain* for each surviving operator. Again, the operator domains for the $\mathbf{J}^{kl}$ are smaller than for the $\mathbf{K}^{kl}$. Since the number of Coulomb and exchange matrices scales linearly with molecular size, and the number of elements per matrix is independent of size, it is evident that the overall number of transformed integrals scales linearly with molecular size.

So far, no approximations were involved by introducing the sparse operator lists and operator domains. However, there are a few terms like

$$\mathbf{G}^{ij} = \ldots - \mathbf{S}\mathbf{T}^{ij} \sum_{kl} \mathbf{L}^{kl}\mathbf{T}^{lk}\mathbf{S} \tag{169}$$

with no coupling between ij and kl via pair amplitudes, and for those terms additional approximations have to be introduced to achieve linear scaling. Fortunately, the integrals involved in these contractions diminish quickly with increasing

distance between the pairs (ij) and (kl), and it is well justified to neglect couplings between remote pairs. For a detailed discussion of these approximations we refer to Ref. [61].

Another important feature of LCCSD is the fact that the number of 3-external and 4-external integrals $(ir|st)$ and $(rs|tu)$ also scales linearly with molecular size, and in fact remains rather modest. For the 3-external integrals this follows from the fact that the r, s in the operators

$$J(\mathbf{E}^{kj})_{rs} = \sum_t (rs|tk)t_t^j \ , \qquad K(\mathbf{E}^{kj})_{rs} = \sum_t (rk|ts)t_t^j \tag{170}$$

are restricted to the J-operator domain $[kj]$, while t in the sum is restricted to the pair domain $[jj]$, which is identical to the orbital domain $[j]$. For the 4-external integrals the PAO indices simply all belong to the same pair domain $[ij]$, since there is a one-to-one correspondence between the residual $(\mathbf{R}^{ij})_{rs}$ and the external exchange operators [53,61]

$$K(\mathbf{T}^{ij})_{rs} = \sum_{tu} T_{tu}^{ij}(rt|us) \ . \tag{171}$$

Thus the number of 4-external integrals per pair is a constant. Fig. 5 shows the number 3-external and 4-external integrals in the local basis as a function of the length n of a linear polyglycine peptide chain $(\mathrm{Gly})_n$ in a cc-pVDZ basis. Even for a molecule as large as $(\mathrm{Gly})_{20}$ with about 1500 basis functions and almost 500 correlated electrons, the disk storage requirement to hold the 3-external integrals is less than 1.5 GByte (compared to more than 3000 GByte in the canonical case). A similar amount is required for the 4-external integrals. Disk storage of the 3-external and 4-external integrals is very appealing, since then the computational cost per iteration is minimized. It can be estimated that forming the contractions of the 3-external and 4-external integrals with the amplitudes would take virtually no time (e.g., less than 50 sec for $(\mathrm{Gly})_{20}$). However, the transformation for the 4-external integrals is quite complicated and has not been implemented so far. Alternatively, the contribution of these integrals can be accounted for by computing for each strong pair an external exchange operator, as defined in eq. (107). In an integral direct scheme, as will be discussed in section 6.5, it is then also possible to achieve linear cost scaling.

4.3 Local connected triples corretion

The ultimate bottleneck for accurate conventional coupled cluster calculations is the connected triples correction, as outlined in section 2.7. If canonical orbitals are used, the Fock matrix is diagonal, and the perturbative energy correction can be obtained directly without storing the triples amplitudes. In the local case this is no longer the case, and in principle an iterative scheme is required, as described above for the local MP2. One might therefore think that the evaluation of a local triples correction for large molecules is impossible, since the storage requirements for all triples amplitudes would scale as $\mathcal{O}(\mathcal{N}^6)$. However, as for the doubles,

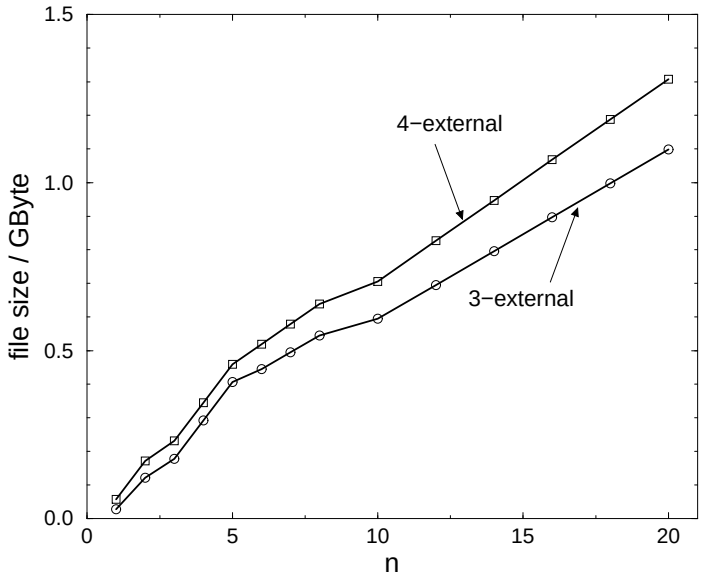


Figure 5. Number of transformed integrals for local CCSD calculations for glycine peptides $(\mathrm{Gly})_n$ as function of the chain length.

triple domains can be introduced, and the correlation of distant electrons can be neglected.

The theory has been outlined in Ref. [62]. Similarly to the LMP2 case the triples amplitudes are obtained by solving a system of linear equations

$$Q_{rst}^{ijk} + W_{rst}^{ijk} = 0 \tag{172}$$

with

$$\begin{aligned} Q_{rst}^{ijk} = &\sum_v \{ \sum_{r's'} f_{tv} T_{r's'v}^{ijk} S_{rr'} S_{ss'} + \text{ permutations} \} \\ &- \sum_m \{ \sum_{r's't'} f_{km} T_{r's't'}^{ijm} S_{rr'} S_{ss'} S_{tt'} + \text{ permutations} \} \end{aligned} \tag{173}$$

and

$$\begin{aligned} W_{rst}^{ijk} = &\sum_v \{ \sum_{r'} (vs|tk) T_{r'v}^{ij} S_{rr'} + \text{ permutations} \} \\ &- \sum_m \{ \sum_{r's'} (mj|kt) T_{r's'}^{im} S_{rr'} S_{ss'} + \text{ permutations} \}. \end{aligned} \tag{174}$$

These equations have to be solved iteratively, and therefore all triples amplitudes T_{rst}^{ijk} must be stored on disk. This seems devastating at a first glance, but by virtue of the local approximations the number of amplitudes can be drastically reduced:

Firstly, a sparse triples list (ijk) of *strong triples* is constructed by restricting the related pairs ij, ik and jk to *strong pairs*. The number of strong triples then scales linearly with molecular size. Secondly, the excitations are restricted to *triple domains* $[ijk]$, constructed as the *union* of the three *strong pair domains*, i.e., $[ijk] = [ij] \cup [ik] \cup [jk]$. Since the sizes of the individual pair domains are independent of the molecular size, the size of the triples domain $[ijk]$ is also independent of the molecular size, yielding overall an asymptotically linear scaling of the number of triples amplitudes.

Another important implication of the constant size of the triple domains is that the number of required 3-external integrals occurring in eq. (174) scales only *linearly* with molecular size. In practice, for each orbital l a *united triple domain* $UT(l)$ is defined as the union of all triple domains $[ijk]$ comprising a common LMO index l, i.e.,

$$UT(l) = \cup[ijk], \quad \text{for} \quad (i = l) \vee (j = l) \vee (k = l), \tag{175}$$

and all 3-external integrals $(vs|tl)$ with $v, s, t \in UT(l)$ are generated using and integral-direct transformation module. Obviously, the size of $UT(l)$ is independent of the molecular size, and the CPU time as well as memory and disk requirements of the transformation scale asymptotically linear with molecular size. In fact, the set of 3-external integrals needed for the triples correction remains pretty small [62], and usually it is a subset of the 3-external integral set required in the preceeding coupled cluster calculation (cf. Fig. 5 in section 4.2).

A linear scaling algorithm for local triples has been implemented in MOLPRO 2000. So far, inter-triples couplings via the occupied-occupied off-diagonal Fock matrix elements are neglected (couplings via the virtual-virtual block and the overlap matrices are included though). This yields about 95% of the local triples correction and has the advantage that the iterative solution of eq. (172) can be avoided. As in the canonical case, the correlation contribution of each individual triple can be computed separately. First test results[62] presented in Table 3 are very promising, showing already for medium sized molecules speedups by factors 500-1000 compared to the conventional (T) calculation presented earlier in Table 1 (note that the calculations in Table 1 used molecular symmetry, while the current calculations were done with no symmetry). In these calculations about 85% of the canonical triples correction was recovered[62]. The savings quickly increase with increasing molecular size. In sharp contrast to the conventional case, the time to compute the local triples corrections is very small as compared to time for the preceding integral transformation and LCCSD calculation. Considering the efficiency of the new triples kernel, it seems even possible to go beyond the CCSD(T) model, i.e. to include the triples into the CC iterations, even for large chemical systems.

Finally, it should be emphasized that the triples amplitudes *can* be stored and an iterative full local triples algorithm is presently under development.

Table 3. CPU, disk and memory requirements for computing the (T) correction. All calculations were performed with a development version of MOLPRO 2000[75]. No molecular symmetry was used.

Molecule	bf	Memory/MW	Disk[a]/MW	CPU/sec[b]
$(Gly)_1$	95	2.57	7.46	187.3
$(Gly)_2$	166	6.38	25.51	757.8
$(Gly)_3$	236	8.82	39.98	934.6
$(Gly)_4$	308	13.28	61.06	1296.6
$(Gly)_6$	450	18.95	94.17	1852.5

a) Disk space for storage of 3-external integrals necessary for (T) only.
b) HP J282 PA8000/180MHz.

5 Multireference electron correlation methods

5.1 Configuration Interaction: general aspects

For a given orbital basis set, Schrödinger's equation as expressed using the second-quantized hamiltonian $\hat{H}$ (equation (34)) is solved by finding eigenvectors and eigenvalues of the hamiltonian matrix in the complete basis of N-electron orbital products. This *full CI* problem is of extremely large dimension for even a small number of electrons with a modest orbital basis size, and is usually intractable. However, it is important to consider it for two reasons: first of all, where the full CI problem can be solved, it provides very important benchmark data against which approximate methods can be evaluated; secondly, the techniques and algorithms applicable to the full CI problem serve as appropriate building blocks for the sometimes more complicated approximate methods.

Although the full configuration space for N electrons in m spatial orbitals consists formally of the complete set of $(2m)^N$ spin-orbital products $\psi_{i_1}(\mathbf{x_1})\psi_{i_2}(\mathbf{x_2})\dots\psi_{i_N}(\mathbf{x_N})$, the space can be reduced substantially through symmetry considerations:

- *Spatial (point group) symmetry.* $\hat{H}$ is invariant to geometrical transformations whose only effect is to interchange identical nuclei. The action of the symmetry operators on the wavefunction is defined through

$$\hat{T}\Psi(q) = \Psi(\hat{T}^{-1}q) \tag{176}$$

  where q represents the coordinates of the particles. In electronic structure calculations, the use of abelian point group symmetry is straightforward; provided each orbital is a basis for an irreducible representation, then so is every orbital product. All orbital products not of the required symmetry can then be simply discarded from the basis. For non–abelian point groups, orbital products are in general of mixed symmetry, and it is therefore usual to exploit the symmetry of only the highest abelian subgroup.

- *Permutational symmetry.* $\hat{H}$ is totally symmetric in the labels of the electrons, and so is invariant under the operation $\hat{I}_{ij}$ which interchanges the labels of

electrons i, j, i.e., $[\hat{H}, \hat{I}_{ij}] = 0$. At the simplest level, $\hat{I}_{ij}^2 = 1$, and so there are $\frac{1}{2}N(N-1)$ two dimensional symmetry groups $\{1, \hat{I}_{ij}\}$. Symmetry adapted wavefunctions will satisfy $\hat{I}_{ij}\Psi = \pm\Psi$, the different signs corresponding to boson and fermion states. We are interested only in fermion solutions, and so it is vital to use this symmetry to exclude unwanted boson and non-physical states. In further detail, there is actually a total of $N!$ *permutations* of the electron labels, which can be build as products of $\hat{I}_{ij}$ operators. As with point groups, we define the action of a permutation operator on the wavefunction through equation (176). The permutations form a group isomorphic with the *Symmetric Group* S_N, and to use permutational symmetry to the full, we must consider all of these $N!$ operators which commute with $\hat{H}$. Since the electronic wavefunction is antisymmetric with respect to all the $\hat{I}_{ij}$, it must form a basis for the one dimensional totally antisymmetric representation of S_N; the representation matrix elements $\Gamma(\hat{P})$ are equal to the *parity* ϵ_P of the permutation $\hat{P}$, which is ± 1 according to whether $\hat{P}$ is made up from an even or odd number of interchanges $\hat{I}_{ij}$. To enforce the symmetry, we apply a multiple of the Wigner projection operator for this representation, the *antisymmetrizer*

$$\hat{\mathcal{A}} = \frac{1}{\sqrt{N!}} \sum_{P}^{N!} \epsilon_P \hat{P} \, . \tag{177}$$

When applied to a simple product of orbitals, $\hat{\mathcal{A}}$ yields the corresponding Slater determinant

$$\begin{aligned} \hat{\mathcal{A}}\,\psi_1(1)\psi_2(2)\ldots\psi_N(N) &= \frac{1}{\sqrt{N!}} \sum_{P}^{N!} \epsilon_P \hat{P}\,\psi_1(1)\psi_2(2)\ldots\psi_N(N) \\ &= \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(1) & \psi_2(1) & \ldots & \psi_N(1) \\ \psi_1(2) & \psi_2(2) & \ldots & \psi_N(2) \\ \vdots & \vdots & & \vdots \\ \psi_1(N) & \psi_2(N) & \ldots & \psi_N(N) \end{vmatrix} . \end{aligned} \tag{178}$$

Note that, apart from a possible phase factor, exactly the same determinant would arise if $\hat{\mathcal{A}}$ were applied to a string of the same orbitals, but in a different order, e.g., $\psi_2(1)\psi_3(2)\psi_1(3)\ldots$. Therefore we can symmetry reduce the full set of m^N orbital products to a much smaller basis of $\binom{m}{N}$ Slater determinants obtained by acting with the antisymmetrizer on each of the $\binom{m}{N}$ unique orbital products. The valid unique orbital products can be determined by assuming an ordering for the orbitals; each of the m orbitals ψ_i is assigned a sequence number i, $i = 1, 2, \ldots, m$, and only orbital products $\psi_{i_1}\psi_{i_2}\ldots\psi_{i_N}$ for which $i_1 < i_2 < \ldots < i_N$ are included.

- *Spin symmetry.* The electron spin operators are defined through

$$\hat{S}^2 = \hat{\mathbf{S}} \cdot \hat{\mathbf{S}}; \;\; \hat{\mathbf{S}} = \sum_i^N \hat{\mathbf{s}}(i)$$
$$\hat{s}_x\alpha = \frac{1}{2}\beta; \;\; \hat{s}_y\alpha = -\frac{i}{2}\beta; \;\; \hat{s}_z\alpha = \frac{1}{2}\alpha;$$
$$\hat{s}_x\beta = \frac{1}{2}\alpha; \quad \hat{s}_y\beta = \frac{i}{2}\alpha; \quad \hat{s}_z\beta = -\frac{1}{2}\beta, \tag{179}$$

where α and β are the one electron spin eigenfunctions. The non–relativistic hamiltonian contains no spin operators, and so $[\hat{H}, \hat{S}_z] = [\hat{H}, \hat{S}^2] = 0$. It is not possible to use simple group theory to exploit these symmetries, since the operators $\hat{S}_z$, $\hat{S}^2$ do not form a closed finite group. But we can use other considerations to force the N electron basis set, and hence the wavefunction, to be eigenfunctions of $\hat{S}_z$ and/or $\hat{S}^2$.

In the case of $\hat{S}_z$, the approach which is usually used is to use a basis of $2m$ orbitals, made up of m spatial orbitals $\phi_i, i = 1, 2, \ldots m$, each multiplied by a spin function α or β. Then any orbital product, or Slater determinant, is automatically an eigenfunction of $\hat{S}_z$ according to (179), with eigenvalue $\frac{1}{2}(N_\alpha - N_\beta)$, where N_α is the number of α–spin orbitals ϕ_i^α in the function, and $N_\beta = N - N_\alpha$. Thus the basis is already adapted to $\hat{S}_z$ symmetry, and we may discard all those N electron functions with the wrong $\hat{S}_z$ eigenfunction. This reduces the size of the Slater determinant basis from $\binom{2m}{N}$ to

$$M_D = \binom{m}{N_\alpha}\binom{m}{N_\beta}, \tag{180}$$

since for each of the $\binom{m}{N_\alpha}$ possible arrangements of the α spin orbitals there are $\binom{m}{N_\beta}$ choices for the β–spin orbitals.

For $\hat{S}^2$, the situation is not so simple. Orbital products or Slater determinants are not in general eigenfunctions of $\hat{S}^2$; for example, following (179), $\hat{S}^2\phi_1^\alpha(1)\phi_2^\beta(2) = \phi_1^\alpha(1)\phi_2^\beta(2) + \phi_1^\beta(1)\phi_2^\alpha(2)$. If the symmetry is to be exploited, Slater determinants must be linearly combined into functions which are eigenfunctions of $\hat{S}^2$. Such functions are often termed *Configuration State Functions* (CSFs). As a simple example, for two electrons in two orbitals with $N_\alpha = N_\beta = 1$, the normalized Slater determinants are

$$\hat{\mathcal{A}}\phi_1^\alpha\phi_1^\beta, \qquad \hat{\mathcal{A}}\phi_1^\alpha\phi_2^\beta, \qquad \hat{\mathcal{A}}\phi_2^\alpha\phi_1^\beta, \qquad \hat{\mathcal{A}}\phi_2^\alpha\phi_2^\beta;$$

the normalized CSFs with $S = 0$ are

$$\hat{\mathcal{A}}\phi_1^\alpha\phi_1^\beta, \qquad \hat{\mathcal{A}}\phi_2^\alpha\phi_2^\beta, \qquad (1/\sqrt{2})(\hat{\mathcal{A}}\phi_1^\alpha\phi_2^\beta + \hat{\mathcal{A}}\phi_2^\alpha\phi_1^\beta),$$

and the CSF with $S = 1$ (i.e., the eigenvalue of $\hat{S}^2$ is $S(S+1) = 2$) is

$$(1/\sqrt{2})(\hat{\mathcal{A}}\phi_1^\alpha\phi_2^\beta - \hat{\mathcal{A}}\phi_2^\alpha\phi_1^\beta).$$

Generally, the set of Slater determinants exactly spans the sets of CSFs with spin quantum numbers $S = \frac{1}{2}(N_\alpha - N_\beta), \frac{1}{2}(N_\alpha - N_\beta) + 1, \ldots, \frac{1}{2}N$. Ignoring

any point group symmetry, the number of CSFs with spin quantum number S is given by the Weyl formula [76]

$$M_C = \frac{2S+1}{m+1}\binom{m+1}{\frac{1}{2}N-S}\binom{m+1}{m-\frac{1}{2}N-S}, \tag{181}$$

for the case that $S = \frac{1}{2}(N_\alpha - N_\beta)$. So, for example, for $S = 0$ and large m, the number of CSFs is less than the number of Slater determinants by a factor of about $\frac{1}{2}N + 1$. The advantage of reducing the basis in this way has to be offset against the increased complexity of the functions which must be dealt with; in practice both Slater determinants and CSFs are commonly used, and we discuss the practicalities of matrix element evaluation with each below.

- *Orbital rotation symmetry.* If we have *all* (unique) orbital products possible for N electrons in m orbitals, then the basis is invariant to rotations (or in fact any non–singular linear transformation) of the orbitals amongst themselves. These rotations form a continuous group $U(m)$, the unitary group (or $GL(m)$, the general linear group), and the theory of such groups is exploited to advantage, for example, in the Graphical Unitary Group Approach (GUGA) [77] for configuration interaction.

In order to perform a variational configuration interaction calculation in either the full or a truncated configuration space, it is necessary to find an eigenvector of the matrix $\mathbf{H}$ of the hamiltonian operator $\hat{H}$ in the appropriate configuration space. Direct construction and diagonalization of $\mathbf{H}$ is usually out of the question since it is typically of dimension 10^3–10^7; but algorithms to find a few eigenvectors for such matrices exist[3,78], and rely on the construction, for a few ($\sim$ 10–20) given trial vectors $\mathbf{c}$, of the action of $\mathbf{H}$ on $\mathbf{c}$,

$$\mathbf{v} = \mathbf{H}\mathbf{c} \, . \tag{182}$$

Other ab initio approaches which are not simple matrix eigenproblems can also proceed through (182). Therefore it is vital to have an efficient scheme for constructing (182) from the hamiltonian integrals $h_{pq}, (pq|rs)$. Following (37), this means we must be able to compute rapidly the set of one and two particle coupling coefficients d^{IJ}_{pq}, D^{IJ}_{pqrs}.

In many circumstances, the most efficient schemes for building (182) require computation only of the one particle coefficients d^{IJ}_{pq}, without explicit construction of the two body terms D^{IJ}_{pqrs}. This is achieved through a formal insertion of the resolution of the identity as a sum over the complete space of orbital products,

$$\begin{aligned} D^{IJ}_{pqrs} &= \langle \Phi_I | \hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps} | \Phi_J \rangle \\ &= \sum_K \langle \Phi_I | \hat{E}_{pq} | \Phi_K \rangle \langle \Phi_K | \hat{E}_{rs} | \Phi_J \rangle \; - \; \delta_{qr} \langle \Phi_I \hat{E}_{ps} | \Phi_J \rangle \\ &= \sum_K d^{IK}_{pq} d^{KJ}_{rs} \; - \; \delta_{qr} d^{IJ}_{ps} \, . \end{aligned} \tag{183}$$

Note that $\hat{E}_{pq}$ commutes with electron label permutations and spin operators; therefore the set of intermediate states $\{\Phi_K\}$ can be reduced to the full set of Slater

determinants or CSFs as convenient; but the same is not true for point group operations, and $\{\Phi_K\}$ must therefore extend over all spatial symmetries. The algorithm for building (182) then proceeds as [79]

$$\begin{aligned}
&\text{DO } K = 1, M \\
&\quad \text{DO } p, q = 1, m \text{ such that } d_{pq}^{KJ} \neq 0 \\
&\qquad F_{pq}^{K} = F_{pq}^{K} + d_{pq}^{KJ}\, c_J \\
&\quad \text{END DO} \\
&\text{END DO}
\end{aligned} \tag{184}$$

$$\begin{aligned}
&\text{DO } r \geq s \\
&\quad \text{DO } p \geq q \\
&\qquad \text{DO } K = 1, M \\
&\qquad\quad E_{rs}^{K} = E_{rs}^{K} + F_{pq}^{K}\, (pq|rs) \\
&\qquad \text{END DO} \\
&\quad \text{END DO} \\
&\text{END DO}
\end{aligned} \tag{185}$$

$$\begin{aligned}
&\text{DO } K = 1, M \\
&\quad \text{DO } p, q = 1, m \text{ such that } d_{pq}^{IK} \neq 0 \\
&\qquad v_I = v_I + E_{pq}^{K}\, d_{pq}^{IK} \\
&\quad \text{END DO} \\
&\text{END DO}
\end{aligned} \tag{186}$$

The one electron part and second term of (183) are easily dealt with in an additional stage, or may be included in (184–186) by modifying the two electron integrals. The advantage of using this scheme is that, for sufficiently large cases, the computation time is dominated by (185), requiring approximately $\frac{1}{2}Mm^4$ floating point operations, and this step is a large dimension matrix multiplication capable of driving most computer hardware at optimal speeds. In what follows, therefore, we are concerned principally with the evaluation, rapidly and in the correct order for assembly of (184–186), of the non–zero d_{pq}^{IJ}, without the need to consider the more complicated structure, and much larger number, of D_{pqrs}^{IJ} coefficients. In some circumstances, simple Slater determinants offer the most efficient route to calculating (182), whilst elsewhere the greater compactness of the CSF basis is important. Therefore we develop techniques for evaluating d_{pq}^{IJ} in both types of basis set.

5.2 Matrix elements between Slater Determinants

Any Slater determinant can be written in the form

$$\Phi_{I,J} = \hat{\mathcal{A}}\left({}^{\alpha}\Phi_I\, {}^{\beta}\Phi_J\right) \tag{187}$$

where ${}^{\alpha}\Phi_I(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_\alpha})$ is a *string* (product) of occupied α–spin orbitals

$$ {}^{\alpha}\Phi_I = \phi^{\alpha}_{I_1}(1)\phi^{\alpha}_{I_2}(2)\ldots\phi^{\alpha}_{I_N}(N) \; , \tag{188} $$

which is completely specified by the ordered list of sequence numbers of occupied orbitals, $\{I_1 < I_2 < \ldots < I_N\}$. Similarly, ${}^{\beta}\Phi_J(\mathbf{r}_{N_\alpha+1}, \mathbf{r}_{N_\alpha+2}, \ldots, \mathbf{r}_N)$ is a string of occupied β–spin orbitals. For the case of a complete basis of determinants, this is a particularly helpful classification, since a wavefunction Ψ is then specified by a fully populated rectangular matrix of coefficients $\mathbf{C}$,

$$ \Psi = \sum_{IJ} C_{IJ}\Phi_{I,J} \; , \tag{189} $$

and this simple rectangular addressing structure makes for a particularly efficient computer implementation. For certain special types of incomplete CI expansion, it is possible to obtain similar structures[80], but it is the case of full CI (FCI) for which the determinant basis has found particularly useful application.

For the evaluation of coupling coefficients, we can exploit the fact that the orbital excitation operator partitions as

$$ \hat{E}_{pq} = \hat{e}^{\alpha}_{pq} + \hat{e}^{\beta}_{pq} \; , \tag{190} $$

where $\hat{e}^{\alpha}_{pq}$, $\hat{e}^{\beta}_{pq}$ excite only α, β spin orbitals respectively; thus the effect of $\hat{E}_{pq}$ on any determinant is to produce at most two new determinants:

$$ \hat{E}_{pq}\hat{\mathcal{A}}\left({}^{\alpha}\Phi_I\,{}^{\beta}\Phi_J\right) = \hat{\mathcal{A}}\left(\left(\hat{e}^{\alpha}_{pq}\,{}^{\alpha}\Phi_I\right){}^{\beta}\Phi_J\right) + \hat{\mathcal{A}}\left({}^{\alpha}\Phi_I\left(\hat{e}^{\beta}_{pq}\,{}^{\beta}\Phi_J\right)\right) \; . \tag{191} $$

Note that the excitation $\hat{e}^{\alpha}_{pq}\,{}^{\alpha}\Phi_I$ is completely independent of ${}^{\beta}\Phi_J$, and so once a particular α–spin excitation has been characterized, one can use the information found for all β strings, obtaining

$$ \langle\Phi_{I,J}|\hat{E}_{pq}|\Phi_{K,L}\rangle = \langle{}^{\alpha}\Phi_I|\hat{\mathcal{A}}\hat{e}^{\alpha}_{pq}\hat{\mathcal{A}}|{}^{\alpha}\Phi_K\rangle\,\delta_{JL} \; . \tag{192} $$

For this to be non zero, ${}^{\alpha}\Phi_I$ must be identical to ${}^{\alpha}\Phi_K$ apart from the replacement of ϕ^{α}_q by ϕ^{α}_p. Suppose that in Φ_I, ϕ_p appears as a function of electron i, and in Φ_K, ϕ_q is correspondingly in position j, i.e.,

$$ {}^{\alpha}\Phi_I = \phi_{I_1}(1)\phi_{I_2}(2)\ldots\phi_{I_{i-1}}(i-1)\phi_p(i)\phi_{I_{i+1}}(i+1)\ldots\phi_{I_j}(j)\ldots \tag{193} $$

and

$$ {}^{\alpha}\Phi_K = \phi_{I_1}(1)\phi_{I_2}(2)\ldots\phi_{I_{i-1}}(i-1)\phi_{I_{i+1}}(i)\phi_{I_{i+2}}(i+1)\ldots\phi_q(j)\ldots \tag{194} $$

Then

$$ \hat{E}_{pq}\,{}^{\alpha}\Phi_K = \phi_{I_1}(1)\phi_{I_2}(2)\ldots\phi_{I_{i-1}}(i-1)\phi_{I_{i+1}}(i)\phi_{I_{i+2}}(i+1)\ldots\phi_p(j)\ldots \tag{195} $$

This is not the same as the string ${}^{\alpha}\Phi_I$, but is related to it by a permutation of the electron labels, known as the *line–up permutation* $\hat{L}$, which in this case is the *cyclic permutation* $\hat{C}(i,j)$, defined through

$$ \hat{C}(i,j)\phi_1(i)\phi_2(i+1)\ldots\phi_{j-i+1}(j) = \phi_{j-i+1}(i)\phi_1(i+1)\ldots\phi_{j-i}(j) \; ; \tag{196} $$

Thus $\hat{L}\hat{e}^{\alpha}_{pq}\,{}^{\alpha}\Phi_K = \hat{C}(i,j)\hat{e}^{\alpha}_{pq}\,{}^{\alpha}\Phi_K = {}^{\alpha}\Phi_I$. For any permutation $\hat{P}$, the following is true:

$$ \hat{P}\hat{\mathcal{A}} = \hat{\mathcal{A}}\hat{P} = \epsilon_P\hat{\mathcal{A}} \; , \tag{197} $$

and so the matrix element (192) is

$$\begin{aligned}\langle {}^{\alpha}\Phi_I|\hat{\mathcal{A}}\hat{e}^{\alpha}_{tu}\hat{\mathcal{A}}|{}^{\alpha}\Phi_K\rangle &= \langle {}^{\alpha}\Phi_I|\hat{\mathcal{A}}\hat{L}^{-1}\hat{\mathcal{A}}|{}^{\alpha}\Phi_I\rangle \\ &= \epsilon_L\sqrt{N!}\,\langle {}^{\alpha}\Phi_I|\hat{\mathcal{A}}|{}^{\alpha}\Phi_I\rangle \\ &= \epsilon_L\ , \end{aligned} \tag{198}$$

since $\hat{\mathcal{A}}^2 = \sqrt{N!}\hat{\mathcal{A}}$, and the only non–zero contribution to $\langle {}^{\alpha}\Phi_I|\hat{\mathcal{A}}|{}^{\alpha}\Phi_I\rangle$ comes from the identity permutation. Therefore all coupling coefficients are 0 or ± 1, and the sign is determined by the parity of the line–up permutation $\hat{L}$. Hence the construction of $\mathbf{F}$ in (184) proceeds as

```
DO ^αΦ_K
    DO p,q = 1,m such that ^αΦ_I = ±Ê^α_pq ^αΦ_K exists
        Determine parity ε_L of line–up permutation L̂
        DO ^βΦ_J
            F(K,J,pq) ← ε_L C(I,J)
        END DO
    END DO
END DO
```
(199)

The innermost loop over ${}^{\beta}\Phi_J$ contains no logic or even multiplication and vectorizes perfectly on all pipeline computers. A similar loop structure is required for the contributions from $\hat{e}^{\beta}_{pq}$, and the logic of (186) can be treated in a similar fashion. Because the number of α, β strings is rather small ($\sqrt{M_D}$), all the necessary single excitation information can be computed once and held in high speed storage. The result is a perfectly vectorized, disk free algorithm[81,82], where for reasonably sized problems at least, there is practically no overhead above the cost of the matrix multiplication (185).

There have been a number of algorithmic developments which have further enhanced the efficiency and applicability of the determinant FCI method. Olsen *et al.* [80] showed how it was possible to reduce the operation count to be proportional to N^2m^2 rather than m^4, with, however, some degradation of the vector performance; their method is particularly useful when the ration m/N is relatively large. Zarrabian *et al.* [83] have used an alternative resolution of the identity to (183), with an intermediate summation over $N-2$ electron (rather than N electron) Slater determinants. Again, when m/N is large, there are many fewer of these, allowing for considerable enhancement in efficiency.

5.3 Matrix elements between Configuration State Functions

In order to build a basis of spin–adapted CSFs, we begin by finding explicit *spin functions* Θ, which are not dependent on space coordinates, and which satisfy $\hat{S}^2\Theta = S(S+1)\Theta$. Having done this, we then attempt to build fully symmetry adapted space–spin functions. For a single electron, there are two possible spin functions $\theta(s)$, where s represents the spin coordinate, namely the usual α and β. For N electrons, the complete space of spin functions is then spanned exactly by the

N electron *primitive spin functions*, written as $[\theta_{i_1}\theta_{i_2}\ldots\theta_{i_N}]$ where the function θ_i of the spin coordinates of each electron in turn may be α or β. There are a total of 2^N such functions, and they are eigenfunctions of $\hat{S}_z$, the eigenvalue M_S being $\frac{1}{2}(N_\alpha - N_\beta)$ where N_α is the number of times α appears in the function, and $N_\beta = N - N_\alpha$; it is then convenient to group them together in sets of those functions sharing the same M_S, the number in each set being $\binom{N}{\frac{1}{2}N+M_S}$.

The primitive spin functions are not in general eigenfunctions of $\hat{S}^2$, and so we seek linear combinations Θ_μ which will be spin eigenfunctions. This is achieved most simply by repeated application of standard angular momentum coupling theory[84,85]. If we have two independent physical systems in each of which we have sets of angular momentum eigenfunctions, $\{|J_1M_1\rangle\}$ and $\{|J_2M_2\rangle\}$, then the members of the set of all products of such wavefunctions are not in general eigenfunctions of the total angular momentum for the combined system. But for a given J_1, J_2 and feasible final quantum numbers J, M, it is possible to find exactly one composite eigenfunction

$$|JM\rangle = \sum_{M_1M_2} \langle J_1J_2M_1M_2|JM\rangle |J_1M_1\rangle|J_2M_2\rangle \tag{200}$$

where the number $\langle J_1J_2M_1M_2|JM\rangle$ is a standard *Clebsch–Gordon coefficient*. Note that all the different M_1 and M_2 components appear in the sum, but only a single J_1 and J_2 value is involved. For N electron spin functions, this suggests a recursive scheme whereby N electron functions are made from such a composite of an $N-1$ electron system with a further single electron. The $N-1$ electron functions arise in the same way from $N-2$ electron spin eigenfunctions, the chain being repeated down to a single particle. For each coupling, the value of J_2 is $\frac{1}{2}$, and so the sum over M_2 extends over two possible values, $\pm\frac{1}{2}$, i.e. a contribution involving α for the last electron and a contribution with β. In this *genealogical* construction, each N electron function is fully described by its parentage — the history of the coupling scheme — which can be visualized as a path on the *branching diagram* shown in Figure 6. Because in the angular momentum coupling one need sum only over the M and not the S quantum numbers, there are in general many independent functions having the same S, M_S, but different ancestry, and we label the functions as $\Theta^N_{S,M,\mu}$ where μ is an index which distinguishes functions with different parentage. The number f^N_S of such functions is indicated at each node on the branching diagram, and one can show inductively that

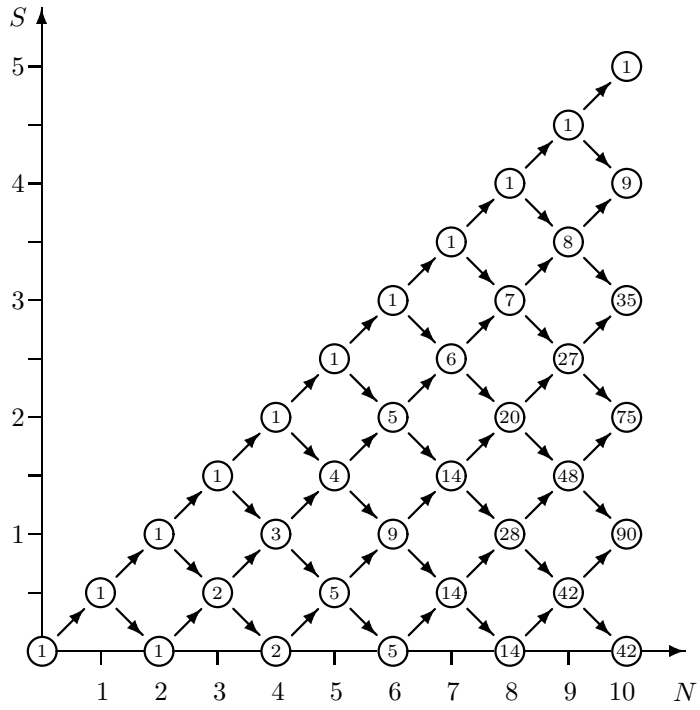
$$f^N_S = \left(\binom{N}{\frac{1}{2}N-S} - \binom{N}{\frac{1}{2}N-S-1}\right) . \tag{201}$$

It follows, again inductively, that

$$\sum_S (2S+1)\, f^N_S = 2^N . \tag{202}$$

It is straightforward to show [86] that the genealogical functions are orthonormal. For each path on the diagram, there are $(2S+1)$ functions, corresponding to the possible different M_S values, and so $\sum_S (2S+1)\, f^N_S$ represents the total number of

Figure 6. The branching diagram



independent N electron branching diagram functions; this is the same as the number of primitive spin functions, and so we have a complete set of spin functions.

Because eventually we need to consider the effect of the antisymmetrizing operator $\hat{\mathcal{A}}$, it is important to develop the permutation properties of the spin functions. Since $\hat{S}^2$ is totally symmetric in the particle labels, it commutes with any permutation, $\hat{P}\hat{S}^2 = \hat{S}^2\hat{P}$. Then it follows that, since $\hat{S}^2\Theta^N_{S,M,\mu} = S(S+1)\Theta^N_{S,M,\mu}$,

$$\hat{S}^2\left(\hat{P}\Theta^N_{S,M,\mu}\right) = \hat{P}\hat{S}^2\Theta^N_{S,M,\mu} \tag{203}$$

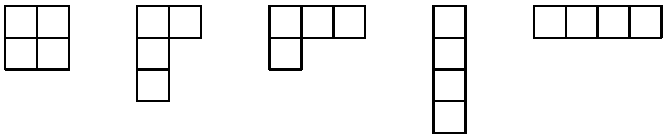
$$= S(S+1)\left(\hat{P}\Theta^N_{S,M,\mu}\right) \ , \tag{204}$$

i.e. $\hat{P}\Theta^N_{S,M,\mu}$ is a spin eigenfunction with quantum numbers S, M. Since $\{\Theta^N_{S,M,\lambda}, \lambda = 1, 2, \ldots, f^N_S\}$ is a complete set of such functions, then $\hat{P}\Theta^N_{S,M,\mu}$ must be a linear combination of these:

$$\hat{P}\Theta^N_{S,M,\mu} = \sum_\lambda \Theta^N_{S,M,\lambda} U_{\lambda\mu}(\hat{P}) \ , \tag{205}$$

i.e., $\{\Theta^N_{S,M,\lambda}, \lambda = 1, 2, \ldots, f^N_S\}$ is a basis for a representation of the symmetric group S_N. The representation is actually isomorphic with particular cases of *Young's Orthogonal Representation*, which is generated (also genealogically) using ideas from

the theory of S_N. Young's orthogonal representation is often depicted graphically. A given representation is drawn as a *Young diagram*, consisting of N adjoining square boxes with rows numbered numerically downwards, and columns rightwards; there may not be more rows in column i than in column $i-1$, nor columns in row j than in row $j-1$. For example, in S_4, the possible Young diagrams are



(206)

For the case of the spin–$\frac{1}{2}$ particles which are our exclusive concern, then only those representations whose Young diagram has at most two rows are relevant, and they correspond to spin quantum numbers S equal to half the difference between the number of boxes in the two rows. Thus for S_4,  ,  ,  represent, respectively, the sets of spin functions with $S = 0,\ 1,\ 2$.

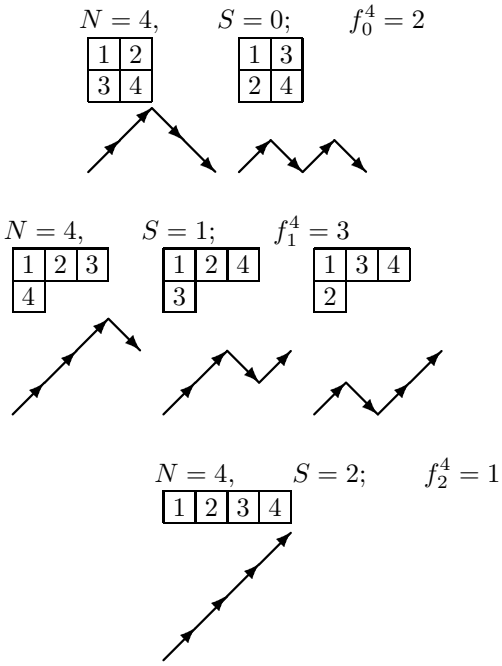
Within each representation, a given basis function is depicted as a *Young tableau*, which is an arrangement of the numbers $1, 2, \ldots, N$ in the Young Diagram, such that numbers always increase along all rows and down all columns. For the two–row Young frames which we consider, the number of such tableaux (i.e., the dimension of the representation) is exactly f_S^N, and in fact there is a one–to–one correspondence between the branching diagram functions and the tableaux; when a particle number appears in the first row, its spin is coupled up, and for those in the second row, the spin is coupled down. For the case of four electrons, the complete set of Young tableaux and corresponding branching diagram functions are shown in Figure 7. The representation matrices $\mathbf{U}(\hat{P})$ constitute all the information which we require for developing properties of the branching diagram functions; for example, the branching diagram functions themselves can be generated from a primitive spin function by use of a suitable projection operator. Formulae for the $U_{\lambda\mu}(\hat{P})$ for any permutation $\hat{P}$ are straightforward to derive from simple rules given in terms of the Young tableaux [86], or, equivalently, from consideration of the Clebsch–Gordon coefficients [86].

Having obtained the representation matrices, we are now in a position to use them in constructing a basis of space and spin functions which are spin eigenfunctions and satisfy the Pauli principle. We write members of this basis as

$$\Phi_{A\lambda} = \hat{\mathcal{A}} \left(\Phi_A \, \Theta^N_{S,M,\lambda} \right) \tag{207}$$

where the spatial function Φ_A is usually an ordered product of spatial orbitals, and $\Theta^N_{S,M,\lambda}$ is a branching diagram function. Note that the antisymmetrizer involves a sum over all permutations $\hat{P}$, and each $\hat{P}$ permutes both the space and the spin

Figure 7. Branching diagram symbols and Young tableaux for 4 electrons



coordinate labels. Inserting the definition of the antisymmetrizer,

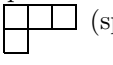
$$\Phi_{A\lambda} = \sqrt{\frac{1}{N!}} \sum_P \epsilon_P \left(\hat{P}_{\rm space}\Phi_A\right)\left(\hat{P}_{\rm spin}\Theta_\lambda\right)$$

$$= \sqrt{\frac{1}{N!}} \sum_P \epsilon_P \left(\hat{P}_{\rm space}\Phi_A\right) \sum_\mu^f U_{\mu\lambda}(\hat{P})\Theta_\mu \tag{208}$$

$$= \sqrt{\frac{1}{f}} \sum_\mu^f \Theta_\mu \Phi_{A\mu\lambda} \ , \tag{209}$$

where we define a set of spatial functions

$$\Phi_{A\mu\lambda} = \sqrt{\frac{f}{N!}} \sum_P \epsilon_P U_{\mu\lambda}(\hat{P})\hat{P}\Phi_A \ . \tag{210}$$

This has the appearance of a projection operator on Φ_A for a representation with matrices $V_{\mu\lambda}(\hat{P}) = \epsilon_P \ U_{\mu\lambda}(\hat{P})$. This is the *conjugate* representation to that supported by the spin functions, and appears in the Young theory as the reversal of the roles of rows and columns, e.g.,  (spin) $\rightarrow$  (space). Note that all

$\Theta_\mu\,, \mu = 1, 2, \ldots, f$ are involved in each of the space–spin functions $\Phi_{A\lambda}$.

For the the coupling coefficients

$$d_{pq}^{A\lambda,B\mu} = \langle \Phi_{A\lambda} | \hat{E}_{pq} | \Phi_{B\mu} \rangle \;, \tag{211}$$

as with determinants, a non–zero contribution will arise only if Φ_A and Φ_B differ by the orbital excitation $\phi_q \rightarrow \phi_p$. Ignoring any complications which arise from doubly occupied orbitals, we must again have $\Phi_A = \hat{L}\hat{E}_{pq}\Phi_B$, where $\hat{L}$ is the appropriate line–up permutation. Inserting (208) into (211) we obtain

$$\begin{aligned} d_{pq}^{A\lambda,B\mu} &= \frac{1}{N!} \sum_{PQ} \epsilon_P \epsilon_Q \langle \hat{P}\Phi_A | \hat{Q}\hat{L}^{-1}\Phi_A \rangle \sum_{\rho\sigma} U_{\rho\lambda}(\hat{P}) U_{\sigma\mu}(\hat{Q}) \langle \Theta_\rho | \Theta_\sigma \rangle \\ &= \frac{1}{N!} \sum_{PQ} \epsilon_P \epsilon_Q \langle \hat{P}\Phi_A | \hat{Q}\hat{L}^{-1}\Phi_A \rangle \sum_{\rho} U_{\rho\lambda}(\hat{P}) U_{\rho\mu}(\hat{Q}) \\ &\qquad\qquad \text{since the spin functions are orthogonal} \\ &= \frac{1}{N!} \sum_{PQ} \epsilon_P \epsilon_Q \langle \hat{P}\Phi_A | \hat{Q}\hat{L}^{-1}\Phi_A \rangle U_{\lambda\mu}(\hat{P}^{-1}\hat{Q}) \;, \end{aligned} \tag{212}$$

using the representation property of $\mathbf{U}(\hat{P})$. Orbital orthogonality then gives the requirement that $\hat{P} = \hat{Q}\hat{L}^{-1}$, and so

$$\begin{aligned} d_{tu}^{A\lambda,B\mu} &= \frac{1}{N!} \sum_{Q} \epsilon_L \epsilon_Q^2 U_{\lambda\mu}(\hat{L}\hat{Q}^{-1}\hat{Q}) \\ &= \epsilon_L U_{\lambda\mu}(\hat{L}) \;. \end{aligned} \tag{213}$$

Thus knowledge of the line–up permutation and the representation matrix elements is sufficient to generate any desired one–particle coupling coefficient.

The above is based on the assumption that ϕ_p and ϕ_q are singly occupied in Φ_A, Φ_B respectively. When one or both orbitals are doubly occupied, further considerations are necessary. Firstly, many of the spin functions give rise to vanishing $\Phi_{A\lambda}$ because of the operation of the Pauli principle acting through the antisymmetrizer. If the orbitals are ordered such that the doubly occupied appear first in their respective pairs, then only those spin functions which couple each pair to singlet are allowed. This of course gives a drastic reduction in the number of possible spin functions, since it is now f_S^N with N referring to the number of singly occupied orbitals only. Following this, there are slight complications to the above scheme for the coupling coefficients; there appear four distinct cases depending on the excited orbital occupancies, of which (213) is one.

How are the relevant representation matrices obtained? Equation (213) shows that one needs all of the representation matrices for all possible cyclic permutations. These matrices can be generated by writing the cycle as a sequence of elementary transpositions,

$$\hat{C}(i,j) = \hat{C}(j-1,j)\hat{C}(j-2,j-1)\ldots\hat{C}(i+1,i+2)\hat{C}(i,i+1) \;; \tag{214}$$

the representation matrices for these transpositions are very sparse, and can be obtained from the shapes of the Young tableaux[86]. The matrix for the cycle is then obtained by matrix multiplication. Unfortunately, this algorithm is too slow for

practical use, and it is much better to precompute and store all of the necessary matrices. The number of matrices that must be stored can be reduced considerably by using a resolution of the identity analogous to that used to factorize two-body matrix elements into sums of products of one-body elements (equation (183)). If we introduce a (fictitious) additional orbital ϕ_α which is defined to occur lexically always after any other orbital, the following identity holds.

$$\hat{E}_{pq} = \hat{E}_{pa}\hat{E}_{aq} \tag{215}$$

This allows one to make use of just those cycles involving the last electron, since the orbital ϕ_a will always be occupied by only this electron in the ordered orbital product string. This is the basis of an efficient algorithm for matrix element evaluation that is fast enough for general use in full and other CI computations[87].

5.4 Molecular Dissociation and the MCSCF method

As discussed in section 1.6, in many situations electron correlation effects are purely of the 'dynamic' type, in the sense that Hartree-Fock is a good zero-order approximation, and under such circumstances, single-reference methods provide an efficient and accurate way to getting correlation energies and correlated wavefunctions. However, wherever bonds are being broken, and for many excited states, the Hartree-Fock determinant does not dominate the wavefunction, and may sometimes be just one of a number of important electronic configurations. If this is the case, single-reference methods, which often depend formally on perturbation arguments for their validity, are inappropriate, and one must seek from the outset to have a first description of the system that is better than Hartree-Fock. Only then can one go on to attempt to recover the remaining dynamic correlation effects.

As in H_2, we can build a general qualitatively correct wavefunction by selecting a number of configurations which are meant to describe all possible dissociation pathways, etc., and then writing the wavefunction as a linear CI expansion

$$\Psi = \sum_I^M c_I \Phi_I \ . \tag{216}$$

The energy is then minimized with respect to not only the c_I (as in the CI method), but also to changes in the common set of orbitals ϕ_t which are used to construct the Φ_I. This orbital optimization is analogous to what is done in the SCF method, hence the name *multiconfiguration self consistent field* (MCSCF), which is given to this approach. Provided all the necessary configurations are included in the set Φ_I, then the method should give a qualitatively correct description of the electronic structure.

Nearly all molecules dissociate to valence states of their constituent atoms, in which only the valence orbitals (e.g., $2s, 2p$ in carbon) are occupied. So ignoring the complications which might occur for Rydberg molecular states, a good description can be obtained by including Φ_I which have only valence orbitals of the molecule occupied. This has important computational consequences, and we distinguish in a calculation the relatively small number of *internal* (or valence) orbitals $\phi_t, \phi_u, \phi_v, \ldots$ from the usually much larger number of *external* orbitals $\phi_a, \phi_b, \ldots$,

which are unoccupied in all configurations, and so actually are not part of the wavefunction. We continue to use the notation $\phi_p, \phi_q, \phi_r, \ldots$ to denote general molecular orbitals from any set. The internal and external orbitals take the roles of the occupied and virtual orbitals in an SCF calculation; as the calculation proceeds, the internal and external orbitals are mixed amongst each other until the optimum internal orbitals are found. Taking these ideas to the extreme suggests the use of a CI expansion consisting of all possible configurations in the valence space, i.e., a FCI type of wavefunction. This approach [88,89,90] is often termed *complete active space SCF* (CASSCF) and has the feature that it is to some extent a 'black box'; the sometimes rather difficult problem of selecting suitable configurations Φ_I is replaced by the simpler identification of important orbitals. If the active orbital space coincides with the true valence space, then correct dissociation at all limits is automatically guaranteed, although there may be many configurations included which are completely unimportant. As a simple example, consider the ground state of N_2. The quartet spin N atom ground state is described by the configuration $2p_x^\alpha 2p_y^\alpha 2p_z^\alpha$. On bringing two N atoms together, one can make 20 CSFs with the correct spin (singlet) and space (A_g in D_{2h}) symmetries, of which one is dominant near equilibrium bond length, but all of which are important at dissociation. The CASSCF wavefunction, a FCI expansion of 6 electrons in 6 orbitals, contains 32 CSFs. Although the ansatz may be wasteful in this way, we note that a complete CI expansion enables the use of special efficient techniques[91], so a CASSCF calculation may actually be easier than a smaller more general MCSCF calculation with the same internal orbital space.

5.5 Determination of MCSCF wavefunctions

We have considered earlier how the matrix elements $H_{IJ} = \langle \Phi_I | \hat{H} | \Phi_J \rangle$ are obtained in terms of one and two electron integrals h_{tu}, $(tu|vw)$ and coupling coefficients d_{tu}^{IJ}, D_{tuvw}^{IJ}:

$$\langle \Phi_I | \hat{H} | \Phi_J \rangle = \sum_{tu} d_{tu}^{IJ} h_{tu} + \tfrac{1}{2} \sum_{tuvw} D_{tuvw}^{IJ} (tu|vw) \; . \tag{217}$$

Thus the expression for the energy is

$$\begin{aligned} E &= \langle \sum_I c_I \Phi_I | \hat{H} | \sum_J c_J \Phi_J \rangle \\ &= \sum_{tu} \sum_{IJ} c_I c_J d_{tu}^{IJ} h_{tu} + \tfrac{1}{2} \sum_{tuvw} \sum_{IJ} c_I c_J D_{tuvw}^{IJ} (tu|vw) \\ &= \sum_{tu} d_{tu} h_{tu} + \tfrac{1}{2} \sum_{tuvw} D_{tuvw} (tu|vw) \; , \end{aligned} \tag{218}$$

where we see the introduction of the one and two electron *density matrices* d_{tu}, D_{tuvw}, which in this context can be viewed as expectation values of the coupling coefficients. This energy expression is the quantity which must be made stationary with respect to changes in the CI coefficients c_I and the orbitals ϕ_t, subject to the

constraints

$$\sum_I c_I^2 = 1 \qquad \text{(normalization)} \tag{219}$$

$$\langle \phi_t | \phi_u \rangle = \delta_{tu} \qquad \text{(orbital orthogonality)} \; . \tag{220}$$

For the CI coefficients, introducing a Lagrange multiplier $\mathcal{E}$ for the first constraint, and setting the differential with respect to c_I to zero, gives the stationary conditions

$$\sum_J \langle \Phi_I | \hat{H} | \Phi_J \rangle c_J - \mathcal{E} c_I = 0 \; , \tag{221}$$

i.e., the usual matrix eigenvalue equations obtained in regular CI theory. For the orbitals, the most straightforward approach is to parametrize orthogonal rotations $\mathbf{U}$ amongst the orbitals ($\phi_t \leftarrow \sum_p \phi_p u_{pt}$) by means of the matrix elements R_{tu} of an antisymmetric matrix. Any orthogonal matrix may be represented as

$$\mathbf{U} = \exp(\mathbf{R}) \qquad \text{where } \mathbf{R}^\dagger = -\mathbf{R} \; . \tag{222}$$

The advantage of this formulation is that the $\frac{1}{2}m(m+1)$ orthogonality constraints are automatically satisfied, leaving $\frac{1}{2}m(m-1)$ free parameters which are contained in the lower triangle of $\mathbf{R}$. There is then no need for Lagrange multipliers, and numerical methods for unconstrained optimization may be used.

To derive the variational conditions for orbital rotations, we note that the orbitals vary on $\mathbf{R}$ through (222) as

$$\left. \frac{\partial \phi_p}{\partial R_{rs}} \right|_{\mathbf{R}=\mathbf{0}} = \delta_{sp} \phi_r - \delta_{rp} \phi_s \; , \tag{223}$$

and that the integrals h_{tu}, $(tu|vw)$ given by (35), (36) are quadratic and quartic, respectively, in the orbitals. Then we obtain

$$\left. \frac{\partial}{\partial R_{rs}} h_{tu} \right|_{\mathbf{R}=\mathbf{0}} = (1 - \tau_{rs})(1 + \tau_{tu}) \delta_{st} h_{ru} \tag{224}$$

$$\left. \frac{\partial}{\partial R_{rs}} (tu|vw) \right|_{\mathbf{R}=\mathbf{0}} = (1 - \tau_{rs})(1 + \tau_{tu})(1 + \tau_{tu,vw}) \delta_{st} (ru|vw) \; , \tag{225}$$

where the operator τ_{ij} permutes the labels i, j in what follows it. Thus the derivative of the energy, which is zero for the converged wavefunction, is given by

$$0 = \frac{\partial E}{\partial R_{rs}} = 2(1 - \tau_{rs}) F_{rs} \; , \tag{226}$$

with

$$F_{rs} = \sum_u d_{su} h_{ru} + \sum_{uvw} D_{suvw} (ru|vw) \; . \tag{227}$$

Equations (221) and (226) must be solved to obtain the MCSCF wavefunction. Note that for some orbital rotations R_{rs}, the variational condition (226) is always obeyed automatically; for example, if both r, s are external, then the density matrix elements are all zero. The same can occur in a more subtle way for certain internal–internal orbital rotations, e.g., for a CASSCF, all internal–internal rotations show this behaviour. When an R_{rs} behaves like this it is known as a redundant variable,

and is best removed from the optimization altogether[92]. Note also that (226) is highly non–linear, in contrast to the linear eigenvalue problem which appears in the CI method; E is 4^{th} order in the orbitals, and infinite order in $\mathbf{R}$, since the orbitals are in fact periodic functions because of the orthogonality constraint.

In order to solve numerically the variational equations (221) and (226), the standard approach is to use some kind of quasi-Newton approach[93,94] that utilizes the gradients of the energy expression to construct a Taylor series for the energy in powers of the parameters that express changes in the wavefunction. Truncation of this power series gives an approximate energy expression that is accurate for small displacements, and which is easier to minimize than the full energy expression. For a given approximate solution, we construct the gradient vector

$$g_\lambda = \left(\frac{\partial E}{\partial p_\lambda} \right)_{\mathbf{p}=\mathbf{0}} \tag{228}$$

and hessian matrix

$$h_{\lambda\mu} = \left(\frac{\partial^2 E}{\partial p_\lambda \partial p_\mu} \right)_{\mathbf{p}=\mathbf{0}} \tag{229}$$

where the set of parameters $\{p_\lambda\}$ contains the changes in CI coefficients $\{\Delta c_I\}$ and the non-redundant orbital change generators $\{R_{rs}\}$. The approximate energy expression

$$E_2(\mathbf{p}) = E_2(\mathbf{0}) + \sum_\lambda g_\lambda p_\lambda + \tfrac{1}{2} \sum_{\lambda\mu} h_{\lambda\mu} p_\lambda p_\mu \tag{230}$$

is then minimized by solving the linear equations

$$0 = g_\lambda + \sum_\mu h_{\lambda\mu} p_\mu \tag{231}$$

The solution $\mathbf{p}$ defines a step that is applied to the wavefunction to improve it. Thus the overall procedure is iterative, each iteration consisting of the construction of the energy, gradient and hessian, followed by solution of the linear Newton-Raphson equations. The Newton-Raphson equations can be very large in dimension, particularly for a large CASSCF full CI expansion; therefore, usually, they have to be solved iteratively as well, using relaxation or expansion vector techniques[95] similar to the Davidson diagonalization algorithm[3]. These iterations are usually referred to as *microiterations* to distinguish them from the enclosing *macroiterations* in each of which a new expansion point is defined.

The generic Newton-Raphson algorithm suffers in this context from two distinct problems associated with robustness and efficiency. First of all, the second-order expansion (230) is valid only for small displacements $\mathbf{q}$, and it is often the case that the predicted step length is outside the 'trust region' of the truncated Taylor series. Modifications that restrict the step length[96], or recast the linear equation system as an eigenvalue problem such that the step length is automatically restricted (*augmented hessian* method[97]) are helpful in improving global convergence. Secondly, however, even with such methods, as many as 20 macroiterations may be required, and each macroiteration is expensive. For each new set of orbitals, in order to construct the gradient and hessian, a subset of the molecular-orbital electron-repulsion

integrals must be constructed, specifically those with up to two external indices ($\mathbf{J}^{tu}, \mathbf{K}^{tu}$), by a computationally demanding transformation of the atomic-orbital integrals, which themselves have to be read from disk or computed on the fly. It is therefore highly desirable to reduce the number of macroiterations. Both problems are solved by adopting an ansatz[98,91] in which the microiterations involve optimization of an approximate energy functional that is second order in the orbital changes themselves, $\Delta\mathbf{T} = \mathbf{U} - \mathbf{1}$, rather than in the generators $\mathbf{R}$. This energy functional is periodic in the orbitals, just like the true energy, and its use gives an algorithm that is much more robust; in fact, in almost all cases, quadratic convergence is seen from the outset, and typically only three macroiterations are needed. Of course, there is additional complication in that the microiterations are solving non-linear rather than linear equations, but these can be effectively addressed using convergence accelerators such as DIIS[99].

5.6 Multireference Perturbation Theory

In order to go beyond a qualitatively correct MCSCF wavefunction Ψ^{REF} and recover as much of the correlation energy as possible, as in the single-reference case, we begin by writing the exact wavefunction in a perturbation series

$$\Psi_{\mathrm{Exact}} = \Psi^{\mathrm{REF}} + \lambda\Psi^{(1)} + \lambda^2\Psi^{(2)} + \dots \,, \tag{232}$$

where λ is an ordering parameter which will eventually be set to 1. Suppose that we can find an operator $\hat{H}^{(0)}$ such that $\hat{H}^{(0)}\Psi^{\mathrm{REF}} = E^{(0)}\Psi^{\mathrm{REF}}$. In the particular case where Ψ^{REF} is the solution of the SCF equations, an appropriate $\hat{H}^{(0)}$ is the many-electron Fock operator,

$$\hat{H}^{(0)} = \sum_{i}^{N} \hat{f}(i) = \sum_{tu}^{m} f_{tu}\hat{E}_{tu} \tag{233}$$

where $\hat{f}$ is the orbital Fock operator; in other cases it may or may not be possible to find a suitable operator, but the arguments we develop still hold. If we write $\hat{H} = \hat{H}^{(0)} + \lambda\hat{H}^{(1)}$, and separate terms of different order in λ in the Schrödinger equation, at first order we obtain

$$\left(\hat{H}^{(0)} - E^{(0)}\right)\Psi^{(1)} + \hat{H}^{(1)}\Psi^{\mathrm{REF}} - \langle\Psi^{\mathrm{REF}}|\hat{H}^{(1)}|\Psi^{\mathrm{REF}}\rangle\Psi^{\mathrm{REF}} = 0 \,. \tag{234}$$

We expand $\Psi^{(1)}$, the first order correction to the wavefunction, and also $\hat{H}\Psi^{\mathrm{REF}}$, the action of the full hamiltonian on the approximate wavefunction, as linear combinations of N-electron configurations in the full space,

$$\begin{aligned} \Psi^{(1)} &= \sum_{I} \Phi_I c_I^{(1)} \\ \hat{H}\Psi^{\mathrm{REF}} &= \sum_{I} \Phi_I h_I \,, \end{aligned} \tag{235}$$

and assume (although again this is not critical) that $\hat{H}^{(0)}\Phi_I = \mathcal{E}_I\Phi_I$. This will be true for the Fock $\hat{H}^{(0)}$ ($\mathcal{E}_I$ is then the sum of the Fock eigenvalues for the orbitals

occupied in Φ_I), and approximately true for others. The first order equation then becomes

$$\sum_I c_I^{(1)} \Phi_I \left(\mathcal{E}_I - E^{(0)} \right) = -\sum_I h_I \Phi_I + \langle \Psi^{\mathrm{REF}} | \hat{H} | \Psi^{\mathrm{REF}} \rangle \Psi^{\mathrm{REF}} \ . \tag{236}$$

This tells us that the basis functions which are required for $\Psi^{(1)}$ are exactly those which appear in the action of $\hat{H}$ on Ψ^{REF}. This set of functions is the *first order interacting space*. Recall that the hamiltonian consists of single and double excitation operators; this means that in turn the first order space consists of all those configurations which are at most doubly excited with respect to the reference function Ψ^{REF}. In the language of second quantization, the first-order space consists of all the non-null configurations $\{\hat{E}_{tu,vw}\Psi^{\mathrm{REF}}\}$.

These arguments can be generalized to higher orders of perturbation theory; at second order, configurations related to the first-order wavefunction by up to double excitations will be introduced, and so the second-order interacting space consists of configurations which are singly, doubly, triply and quadruply excited relative to Ψ^{REF}.

One route to carry these ideas forward is to simply apply regular Rayleigh-Schrödinger perturbation theory to obtain the perturbation series for the energy. With the choice of Fock $\hat{H}^{(0)}$, this is the single-reference Møller-Plesset theory (MP)[100,101] or Many-Body Perturbation Theory (MBPT)[102]. For multiconfigurational Ψ^{REF}, the choice of zero order hamiltonian is not so obviously unique, but a number of different variants have been very successfully used[103,104,105,106,107]. These are generally non-diagonal in the configuration basis, and so solution of the first-order equations must be carried out iteratively; in contrast, for a Hartree-Fock reference with canonical molecular orbitals, each Slater determinant is an eigenfunction of $\hat{H}^{(0)}$, and so the first-order equations have an explicit analytic solution.

Multireference perturbation theory at second order (MRPT2 or CASPT2) is now well established as a robust and reliable technique particularly, for example, in the computation of electronic excitation energies[106], and is computationally feasible in almost all cases where the underlying MCSCF or CASSCF calculation is possible. Third-order perturbation theory[103,108] can also be carried out for smaller systems, and the results show significant differences from second order, indicating the need for caution in the use of CASPT2.

5.7 Multireference Configuration Interaction

Although perturbation theory may be a dangerous tool to rely on, the interacting space hierarchy concept provides useful insight on how to design other methods. If we consider doing a variational CI calculation, we now know that, even though FCI may be impossible, we expect to obtain most of the correlation energy using a basis consisting of the first-order interacting space. In the case of an RHF reference wavefunction Ψ^{REF} this is the singles and doubles (CISD) method, with the basis consisting of all Slater determinants which are related to Ψ^{REF} by a single or double spin-orbital excitation. Strictly speaking, for RHF Ψ^{REF}, singles do not formally enter until second order perturbation theory, but in practice their effect can be quite

significant, and there are fewer of them than doubles, and so they are invariably included as well.

The same kind of approach can be taken for an MCSCF $\Psi^{\rm REF}$. The first-order space is certainly spanned by a wavefunction of the form

$$\Psi = \sum_I c_I \Phi_I + \sum_{Sa} c_a^S \Phi_S^a + \sum_{Pab} C_{ab}^P \Phi_P^{ab} \ , \tag{237}$$

where the three types of configuration Φ_I, Φ_S^a, Φ_P^{ab} contain respectively 0, 1, 2 occupied external orbitals, and the set of configurations is the union of the sets of CSFs obtained by making all possible single and double excitations on each reference configuration in turn. For the case that $\Psi^{\rm REF}$ consists of a single closed shell configuration, (237) is the single-reference CISD wavefunction; when $\Psi^{\rm REF}$ contains more than one configuration, variational treatment of (237) is usually referred to as *multireference CI* (MRCI)[109,110,111,112].

Since there are usually many more external orbitals than internal orbitals, the doubly external configurations Φ_P^{ab} are expected to be by far the most numerous, just as in the single-reference case, and we focus attention on these in considering what work has to be done in evaluating hamiltonian interactions. In the general multi-reference case, it is not possible to arrive at explicit matrix-oriented expressions for the hamiltonian matrix elements. However, some simplification beyond the general CI matrix element strategy presented in section 5.3 is certainly possible; just as in the single-reference case, there is special structure associated with the pairs of external orbitals ϕ_a, ϕ_b. In the formation of CSFs Φ_P^{ab}, it is advantageous to take the occupied orbital string which is inserted into equation (207) such that the orbitals ϕ_a and ϕ_b appear as functions of the coordinates of electrons 1 and 2 respectively; this means that the function is pure singlet or triplet coupled in the two external orbitals, exactly as in the single-reference case, and allows for some simplification in matrix element evaluation. The structure of the wavefunction in the external orbitals is then no more complicated than in the single-reference problem, and so closed formulae for those parts involving external orbitals are obtainable; for example, the contribution from all external integrals has exactly the same form as in single-reference SDCI, and can be obtained efficiently by computing the external exchange matrices for each pair P. However, for the internal orbitals, the CSFs are completely general in character, and ultimately one must compute one and two particle coupling coefficients using the general techniques of section 5.3. For example, that part of the hamiltonian containing the Coulomb integrals, $\sum_{tuab} J_{ab}^{tu} \hat{E}_{ab} \hat{E}_{tu}$, gives rise to matrix elements

$$\langle \Phi_P^{ab} | \hat{H} | \Phi_Q^{cd} \rangle = \frac{1}{2} \delta_{pq} \sum_{mn} \alpha_{mn}(P,Q) \, (1 + p\tau_{ab})(1 + q\tau_{cd}) \, \delta_{bd} \, J_{ac}^{tu} \ , \tag{238}$$

where $p = \pm 1$ according to whether Φ_P^{ab} is singlet or triplet coupled in the external space. $\alpha_{tu}(P,Q)$ is simply a one particle coupling coefficient for the operator $\hat{E}_{tu}$ between the functions Φ_P^{ab} and Φ_Q^{ab},

$$\alpha_{tu}(P,Q) = \langle \Phi_P^{ab} | \hat{E}_{tu} | \Phi_Q^{cd} \rangle \ . \tag{239}$$

Although coupling coefficient evaluation is required, all the coupling coefficients are completely independent of the external orbital labels; thus many hamiltonian matrix elements share the same coupling coefficients in a regular manner. Discovery of this property [113,114] first opened the way for large scale MRCI calculations. Although the coupling coefficient evaluation problem is dramatically reduced by exploiting these special properties, the MRCI method is still severely restricted by computational difficulties. For even quite modest numbers of reference configurations, the number of pair functions Φ_P^{ab} can be rather large; this means that the dimension of the hamiltonian matrix can easily exceed the length of vector which can be stored on the computer, and, more importantly, the number of matrix elements which must be evaluated becomes completely unmanageable. Nevertheless, benchmark calculations, in which MRCI results are compared with those from full CI in the same basis, indicate that MRCI is the ab initio method of choice for all circumstances in which single determinant descriptions do not work, and that very high accuracy may be obtained[115,116].

An alternative formulation which avoids the rapid increase in basis size with the number of reference configurations is possible [113]. Instead of selecting singly and doubly excited CSFs from each reference configuration, we can construct configurations by applying excitation operators to the reference wavefunction as a single entity:

$$\Psi = \sum_{tuvw} C^{tuvw} \hat{E}_{tu,vw} \Psi^{\rm REF} + \sum_{tuva} C_a^{tuv} \hat{E}_{at,uv} \Psi^{\rm REF} + \sum_p \sum_{ab} \sum_{t \geq u} C_{ab}^{tup} \, \tfrac{1}{2} \left(\hat{E}_{at,bu} + p \hat{E}_{au,bt} \right) \Psi^{\rm REF} \; . \tag{240}$$

This is the *internally contracted* MRCI (ICMRCI)[113,117,118] wavefunction, and it is obvious that the number of configurations is now independent of the number of reference functions, depending only on the numbers of internal and external orbitals. In this way, the size of CI expansion is reduced typically by one or two orders of magnitude; the configuration set, however, still spans the first order interacting space, and although CMRCI can be considered as only an approximation to MRCI, benchmark calculations show that in most cases the extra error introduced by the contraction is several times smaller than the error of MRCI relative to full CI[118]. The price that is paid is that the configurations are now much more complicated, being in fact linear contractions of CSFs according to the values of the reference coefficients. This means that coupling coefficient evaluation is now a formidable problem; the simple CSF coupling coefficients are replaced by reduced density matrices of high order. For example, for the Coulomb integrals considered previously, the coupling coefficients are

$$\alpha_{tu}(vwp, xyq) = \delta_{pq}(1 + p\tau_{xy}) \, \langle \Psi^{\rm REF} | \hat{E}_{vx,wy,tu} | \Psi^{\rm REF} \rangle \; . \tag{241}$$

This third-order density matrix is evaluated using the general resolution-of-identity techniques used in the full CI problem, i.e.,.

$$\langle \Psi^{\rm REF} | \hat{E}_{vx,wy,tu} | \Psi^{\rm REF} \rangle = \sum_K \langle \Psi^{\rm REF} | \hat{E}_{vx,wy} | \Phi_K \rangle \langle \Phi_K | \hat{E}_{tu} | \Psi^{\rm REF} \rangle + \text{ lower order terms} \tag{242}$$

where the $\{\Phi_K\}$ are appropriate CSFs. For a given bra (vw) and ket (xy), all the matrix elements $\langle\Psi^{\rm REF}|\hat{E}_{vx,wy}|\Phi_K\rangle$ are found by successively applying the operators $\hat{E}_{av}, \hat{E}_{xa}, \hat{E}_{aw}, \hat{E}_{ya}$ (ϕ_a is a 'fictitious' unoccupied orbital) to $\Psi^{\rm REF}$. For processing a given Coulomb matrix $\mathbf{J}^{tu}$, these matrix elements are combined with precomputed $\langle\Phi_K|\hat{E}_{tu}|\Psi^{\rm REF}\rangle$.

An additional complication in ICMRCI is that the configurations are non–orthogonal in a non–trivial way, and their orthogonalization can be a computational bottleneck[117]. For this reason, the standard approach to ICMRCI is a hybrid that combines the best features of uncontracted and contracted wavefunctions [118]; contraction is carried out only where it is easiest, and of most benefit, namely for the doubly external configurations, and the all-internal and singly-externals are left uncontracted.

An unfortunate feature of an MRCI calculation is that, just as in the single-reference CISD case, the energy is not an *extensive* function of the number of electrons as it should be. This undesirable feature of any truncated variational CI calculation can to some extent be avoided in MRCI by error cancellation across a potential energy surface; provided, for example, dissociation asymptotes are computed as supermolecules rather than by adding fragment energies, reasonable results can be obtained for dissociation energies. It is also true that the size-consistency errors for MRCI are usually much less than for single-reference CISD, since MRCI already contains some of the important quadruple configurations. However, the effects can never be completely avoided.

One way to view the lack of size-consistency in variational CI is by considering the Rayleigh quotient correlation energy functional itself,

$$\mathcal{E} = \frac{\langle\Psi|\hat{H} - E^{\rm REF}|\Psi\rangle}{\langle\Psi|\Psi\rangle}\,. \tag{243}$$

Suppose Ψ is, for example, restricted to contain double excitation configurations only, and that the coefficient of the reference wavefunction is kept fixed (intermediate normalization, $\langle\Psi|\Psi^{\rm REF}\rangle = 1$). Then the numerator of this expression can be shown to grow linearly with system size $\mathcal{N}$; however, the denominator also grows, but as $1 + \lambda\mathcal{N}$, where λ is a constant. This spoils the proper linear scaling of the correlation energy. In the absence so far of problem-free multireference coupled-cluster approaches, this analysis gives rise to a number of approximate ways to correct for the effects of lack of extensivity. The simplest, the Davidson or '+Q' correction [119,26], involves a straightforward rescaling of the correlation energy by $\langle\Psi|\Psi\rangle$, i.e. replacing the denominator of (243) by 1 once the wavefunction has been determined. More explicitly,

$$\mathcal{E}^{\rm CI+Q} = \frac{1 - c_0^2}{c_0^2}\mathcal{E}^{\rm CI}\,, \tag{244}$$

where c_0^2 is the weight of the reference wavefunction $\Psi^{\rm REF}$ in the final normalized CI wavefunction. Alternative approaches (ACPF[120], AQCC[121]) introduce at the outset a denominator in the energy functional that does not increase with system size. This modified approximate functional is then minimized to determine the wavefunction and energy.

6 Integral-direct methods

Since the first formulation of the LCAO finite basis scheme for molecular Hartree-Fock calculations, computer implementations of this method have traditionally been organised as a two-step process. In the first step all the two-electron repulsion integrals (ERIs) over four contracted Gaussian basis functions are calculated and stored externally on disk, while the second step comprises the iterative solution of the Hartree-Fock Roothaan equations, where in each iteration the integrals from the first step are retrieved from disk and contracted with the present density matrix to form a new Fock matrix. This subdivision of the computational process into the two steps was motivated by the relatively high CPU cost necessary to generate the ERIs using rather complicated analytical recurrence relations, which was clearly dominating a Hartree-Fock calculation. For post Hartree-Fock calculations, which are traditionally formulated using the canonical SCF orbitals from a preceding Hartree-Fock calculation as a basis, an integral transformation of the AO ERIs generated in the first step to the canonical MO basis is required prior to the actual correlated calculation. The computational complexity of such an integral transformation scales with $\mathcal{O}(\mathcal{N}^5)$, where $\mathcal{N}$ is a measure of the molecular size or the number of correlated electrons. It also is quite memory and disk intensive. The amount of disk space required to hold the AO (and MO) ERIs scales as $\mathcal{O}(\mathcal{N}^4)$.

The last several decades have witnessed continuous rapid advances in computer technology, and in fact the progress in CPU technology has been much faster than the development of I/O facilities. Furthermore, much effort has been invested in improving integration techniques. Hence, with the conventional two step procedure one now faces the dilemma of being able to compute large numbers of integrals rapidly, but spending a relatively large amount of time and resources in their storage and retrieval. In fact, the size of chemical systems one can handle today with the conventional method described above is primarily limited by the disk space required to store the AO ERIs, rather than the CPU time required to compute these. Integral-direct methods offer a solution to this problem. The philosophy is to eliminate the $\mathcal{O}(\mathcal{N}^4)$ bottleneck of AO ERI storage altogether by recomputing the ERIs on the fly whenever needed, thus trading disk space and I/O load at the expense of additional CPU time. Integral-direct methods were first used in Hartree-Fock (SCF) theory almost two decades ago ("direct SCF" approach by Almlöf *et al.* [122]), and it constituted a break of a paradigm at that time. These days, direct SCF programs are part of virtually all ab initio program packages used by the community. Since the pioneering direct SCF work integral-direct methods have been extended to electron correlation methods like multiconfigurational SCF [123,124,60], many-body perturbation theory [MBPT(2)] [125,126,127,60], MBPT(2) gradients [128] and coupled cluster methods [129,130,60]. In contrast to the SCF method, where the ERIs over atomic orbitals (AOs) (i.e., the basis functions) are immediately contracted to the Fock matrix in AO basis, and only AO integrals are needed, correlation methods including MCSCF require an AO to MO integral transformation, as discussed above. Hence an intermediate four-indexed quantity (rather than the two-indexed Fock matrix in direct SCF procedures) arises and has to be dealt with. A full 4-index transformation, carried out as four quarter transformations

has a flop count that scales as $\mathcal{O}(m^5)$ with the number of basis functions m, and has $\mathcal{O}(m^4)$ storage requirements. At a first sight the storage requirements for such an integral transformation seem to rule out any integral-direct implementation of a correlated method, since no savings to the conventional method seem to be possible. Fortunately enough, however, most correlation methods can be reformulated in terms of AO ERIs and a reasonably small subset of MO integrals [20]. Such MO integral subsets typically have two indices restricted to the *occupied* orbital space of dimension m_{occ}, which is usually much smaller than m. For example, the computation of the MBPT(2) energy requires only the exchange integrals $(ia|jb)$, while for direct MCSCF and all other correlation methods the Coulomb $(ij|pq)$ and exchange $(ip|jq)$ MO integrals are needed. The disk space necessary to hold such a subset of MO integrals then is $\mathcal{O}(m_{\mathrm{occ}}^2 m^2)$, i.e. for a ratio $m/m_{\mathrm{occ}} \approx 10$ this means savings of a factor of 100 and larger in the storage requirements, compared to the conventional method. In the work by Schütz *et al.* [60] it was demonstrated that for integral-direct implementations of most electron correlation methods (MP2-4(SDQ), CCSD, QCISD, BCCD, MCSCF, MRPT2/3, MRCI) only three integral-direct kernel procedures are necessary. The only exception are methods involving triply or higher excited configurations. Apart from the trivial Fock matrix construction routine these involve a generalized partial integral transformation and a module for the construction of *external exchange operators* which corresponds basically to a two-index contraction of AO ERIs with the doubles amplitude matrices, backtransformed to AO basis, as explained in section 2.4.

Integral-direct methods are especially powerful in the context of local correlation methods [57,58,59,53,54]. Here, additional savings are possible by describing occupied and virtual correlation spaces in terms of localized MOs and projected (non-orthogonal) AOs, respectively, which in turn allows to exploit the short range character of dynamic correlation (asymptotic distance dependence is $\propto r^{-6}$ in insulators). In such a scheme, a hierarchical treatment of different electron pairs is possible, depending on relative distance of the corresponding LMOs. Furthermore, the virtual space spanned by the non-orthogonal projected AOs can be partitioned into domains (cf. section 4). As a result of this, only very small subsets of (transformed) integrals are required even for methods including triply excited configurations, and the number of these integrals scales linearly with the molecular size. This, in turn, opens the path for $\mathcal{O}(\mathcal{N})$ electron correlation methods and hence the treatment of very large molecular systems at a level of very high accuracy.

6.1 The direct SCF method

In the most naive implementation, writing a computer code for a direct SCF scheme comprises little more than just replacing the reading of one- and two-electron integrals in the SCF algorithm by their repeated calculation. However, in order to get an efficient program, it is clear that such a change in the paradigm calls for major restructuring of the code. Since the computation of the two-electron integrals is rather expensive, a direct algorithm should be *integral driven*, i.e. integral evaluation concerns should dictate the order of events. Once an integral has been computed, it should be used to the maximum extent possible, as long as no external

storage is invoked.

Two-electron repulsion integrals (ERIs) are integrals of the following form (assuming real basis functions)

$$(\mu\rho|\nu\sigma) = \int\int \chi_\mu(1)\chi_\rho(1) r_{12}^{-1} \chi_\nu(2)\chi_\sigma(2) dr_1 dr_2, \tag{245}$$

where $\chi_\mu, \chi_\rho, \chi_\nu, \chi_\sigma$ denote contracted Cartesian Gaussians,

$$\chi_\mu = \sum_\alpha c_{\alpha\mu}\bar{\chi}_\alpha(r) = \sum_\alpha c_{\alpha\mu}(\bar{\chi}_\alpha^x(x)\bar{\chi}_\alpha^y(y)\bar{\chi}_\alpha^z(z)), \tag{246}$$

with

$$\bar{\chi}_\alpha^x(x) = (x - x_\alpha)^{k_\alpha} \exp[-a_\alpha(x - x_\alpha)^2], \tag{247}$$

and $\bar{\chi}_\alpha^x(x)\ldots$ symbolize Cartesian components of primitive Gaussians, centred at origins $\mathbf{r}_\alpha = (x_\alpha, y_\alpha, z_\alpha)$. Usually, these centres are taken to be the atoms, but sometimes basis functions are also positioned between atoms. One of the most important reason to choose Gaussians as basis functions is the separability into products of Cartesian components, as indicated in eq. (246). Another equally important reason for the efficacy of a Gaussian basis set is the fact that a two-centre product of Gaussians can be expressed as a short expansion of one-centre Gaussians – *the Gaussian Product Theorem, (GPT)*

$$\begin{aligned} \bar{\chi}_\alpha^x(x)\bar{\chi}_\beta^x(x) &= \sum_{i=0}^{k_\alpha+k_\beta} C_i^{k_\alpha+k_\beta}\phi_{Pi}(x), \qquad \text{with} \\ x_P &= \frac{a_\alpha x_\alpha + a_\beta x_\beta}{a_P}, \\ a_P &= a_\alpha + a_\beta, \\ \phi_{Pi}(x) &= x^i e^{a_P(x-x_P)^2} \ . \end{aligned} \tag{248}$$

For the case of two s-type Gaussians ($k_\alpha = k_\beta = 0$) the single expansion coefficient is

$$C_0^0 = \exp[-(a_\alpha a_\beta/a_P)(\mathbf{r}_\alpha - \mathbf{r}_\beta)^2]. \tag{249}$$

In a geometrical interpretation, the GPT states that the *product* of two Gaussian functions (with arbitrary polynomial factors) can be expressed as a finite sum of new Gaussians, all centred at a single point P, which is located on the line connecting the two original centres $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$.

The ERIs as given in eq. (245) can be evaluated analytically using various methods. At the heart of all these methods lies the GPT and some recurrence relations to shift angular momenta from one function to the other. Here, we will not go into the details; for a recent review we refer to Ref. [131].

From eq. (245) it is immediately evident that the ERIs obey the permutational symmetry relations

$$\begin{aligned} (\mu\rho|\nu\sigma) = (\rho\mu|\nu\sigma) = (\mu\rho|\sigma\nu) = (\rho\mu|\sigma\nu) \\ = (\nu\sigma|\mu\rho) = (\sigma\nu|\mu\rho) = (\nu\sigma|\rho\mu) = (\sigma\nu|\rho\mu) \ . \end{aligned} \tag{250}$$

By exploiting this permutational symmetry the number of integrals that need to be evaluated can be reduced by about a factor of eight. In modern quantum chemical codes the ERIs are usually evaluated over *shell quadruplet batches*. A shell typically comprises all contracted functions of a given centre and given angular momentum. For example, an *s*-shell of a 3s2p1d basis set comprises three functions, a *p*-shell six, and a *d-shell* 5 functions. In order to exploit an integral shell quadruplet batch to its maximum extent, i.e. to make use of the permutational symmetry mentioned above, the code should drive *triangularly* over the shell quadruplets. In the following we will use M, R, N, S as symbols for shells of basis functions, i.e., $\mu \in M, \rho \in R$, etc. A direct Fock builder performs a two-index contraction of each integral batch `(MR|NS)` with the related piece of the density matrix. If it runs over the *minimal integral list* (i.e. exploits the full permutational symmetry of the ERIs), each integral batch contributes to the Fock matrix via two Coulomb and four exchange components, as indicated in the pseudocode below.

```
DO M=1,NShell
  DO R=1,M
    DO N=1,M
      DO S=1,N | R (for N=M)
        compute integral shell quadruplet block (MR|NS)
        compute Coulomb component of Fock matrix:
          f(M,R)=f(M,R)+4*(MR|NS)*d(N,S)
          f(N,S)=f(N,S)+4*(MR|NS)*d(M,R)
        compute exchange component of Fock matrix:
          f(R,N)=f(R,N)-(MR|NS)*d(M,S)
          f(R,S)=f(R,S)-(MR|NS)*d(M,N)
          f(M,N)=f(M,N)-(MR|NS)*d(R,S)
          f(M,S)=f(M,S)-(MR|NS)*d(R,N)
      END DO
    END DO
  END DO
END DO
```

6.2 Integral prescreening

Obviously, the ERI supermatrix is a four-indexed quantity. Therefore, the computational effort to evaluate the ERIs scales nominally as $\mathcal{N}^4$, where $\mathcal{N}$ is a measure for the size of the chemical system (e.g. the number of basis functions for a given basis set). For instance, for a system with 100-200 atoms, involving about 2000 basis functions or more, the ERI supermatrix would comprise $10^{12} - 10^{13}$ integrals. It is clear that even though the algorithms for ERI evaluation have been drastically improved over the last two decades, no code can deal with all these integrals in a routine calculation.

In the integral-direct approach the storage bottleneck is removed by reevaluating ERIs on the fly whenever needed. One is then in the situation that the integral evaluation is the bottleneck. The solution to the problem is not only to generate

the ERIs more efficiently, but to search for algorithms that can avoid the calculation of negligible integrals altogether. Fortunately, the ERI supermatrix is very sparse for extended chemical systems. Consider for a moment an ERI $(\mu\rho|\nu\sigma)$, as given in eq. (245). Since both μ and ρ are Gaussian functions and involve the same electron coordinate r_1, it is immediately clear from eqs. (248) and (249) that the integrand decreases exponentially with the distance between the centres $\mathbf{r}_\alpha - \mathbf{r}_\beta$. The same holds for ν and σ. In fact, also the value of the ERI drops exponentially with the distance between μ and ρ or ν and σ. Unfortunately, the two Gaussian pairs $(\mu\rho)$ and $(\nu\sigma)$ are coupled by the Coulomb interaction $1/r_{12}$, which is long range. Hence, the ERI still might be significant even if $(\mu\rho)$ is far away from $(\nu\sigma)$. Therefore, the number of non-vanishing ERIs scales asymptotically with $\mathcal{N}^2$ rather than with $\mathcal{N}^4$. In a direct SCF scheme the ERIs are reevaluated in each iteration and immediately contracted over two indices with the corresponding density matrix elements. Now, for an extended (but non-periodical) chemical system, the density itself is also sparse (i.e. $D(M,N)$ becomes small if M is distant from N), provided that the HOMO-LUMO gap is large enough (which is usually the case for non-metallic systems). Furthermore, the exchange components of the Fock matrix requires contractions of the ERIs where the first index involves one function of the first Gaussian pair $(\mu\rho)$, while the second index corresponds to one function of the second pair $(\nu\sigma)$. Hence, by virtue of the sparsity of the density matrix, the number of ERIs with non vanishing contributions to the Fock exchange component scales asymptotically linear (i.e. as $\mathcal{O}(\mathcal{N})$) with molecular size. Unfortunately, this is not true for the Coulomb component, where the density connects just functions within each pair. Thus, a straightforward scheme would lead to $\mathcal{O}(\mathcal{N}^2)$ scaling. However, since Coulomb repulsion is a relatively simple (i.e. classical) form of interaction, one can employ multipole expansions[132,133,134,135] for the long range interactions, for which linear scaling with molecular size can be achieved. If then the evaluation of the Coulomb and exchange contributions to the Fock matrix is done separately, an overall linear scaling of the Fock matrix construction in integral-direct SCF calculations can be achieved.[136].

A prerequisite for approaching quadratic or even linear scaling in a direct SCF scheme is a method to estimate the integral values as accurately as possible without actually computing them. This estimate must not be done for each integral or each integral batch individually, since then the test would scale itself with $\mathcal{N}^4$ and become the bottleneck. A strict upper bound for the ERI $(\mu\rho|\nu\sigma)$ can be obtained from the Schwartz inequality [137]

$$|(\mu\rho|\nu\sigma)| \leq Q_{\mu\rho}Q_{\nu\sigma}, \qquad \text{with} \qquad Q_{\mu\rho} = \sqrt{(\mu\rho|\mu\rho)}\,. \tag{251}$$

The $Q_{\mu\rho}$ necessary to compute the Schwartz estimates for the ERIs are just two indexed quantities, and can easily be precomputed outside the the nested loop over shell quadruplet batches. The number of non-negligible such integrals scales linearly with molecular size, and it is possible to evaluate them in a way that the overhead with quadratic scaling is very small. Furthermore, since the ERI prescreening takes place at the level of shell batches, only the maximum values of $Q_{\mu\rho}$ over the respective shells, i.e. the

$$Q_{MR} = \underset{\mu\in M,\rho\in R}{\text{Max}}\, Q_{\mu\rho} \tag{252}$$

are required. The four nested shell loops can now be replaced by two loops over the pairs (MR) and (NS) with non-negligible Q_{MR} and Q_{NS}, respectively, and within these loops the product $Q_{MR}Q_{NS}$ can be tested against a threshold. Formally, this prescreening procedure scales quadratically with molecular size, but the prefactor is very small. A more powerful prescreening scheme has also to take the density matrix into account. As we have seen above, each ERI contributes with two Coulomb and four exchange components to the Fock matrix, and therefore the following test is required

$$\begin{aligned} Q_{MR}Q_{NS}d_{\max} \geq \tau, \quad & \text{with} \\ d_{\max} = \max(4|d_{MR}|, 4|d_{NS}|, |d_{MN}|, |d_{MS}|, |d_{RN}|, |d_{RS}|) \; . & \end{aligned} \tag{253}$$

If the exchange component of the Fock matrix is constructed separately, eq. (253) reduces to

$$Q_{MR}Q_{NS}d_{\max} \geq \tau, \quad \text{with} \quad d_{\max} = \max(|d_{MN}|, |d_{MS}|, |d_{RN}|, |d_{RS}|), \tag{254}$$

leading to an overall linear scaling of shell quadruplets that survive the test, and consequently the number of ERIs that have to be computed.

The efficiency of this prescreening scheme can be enhanced in several ways. First, since ERIs are evaluated batchwise over whole shells, it might be desirable to split off diffuse functions (small exponents) from tight functions (large exponents), and to treat diffuse functions in separate shells. Even though this will increase the total number of shell quadruplets, the actual number of integrals to be computed can be reduced. Second, the effectivity of the prescreening schemes in eqs. (253) and (254) can be enhanced further by constructing *incremental* Fock matrix updates in each new iteration, rather than the total Fock matrix. Consider the the Fock matrices of two consecutive iterations $m-1$ and m:

$$\begin{aligned} f_{\mu\rho}^{(m-1)} &= h_{\mu\rho} + \sum_{\nu\sigma} d_{\nu\sigma}^{(m-1)}\{2(\mu\rho|\nu\sigma) - (\mu\nu|\rho\sigma)\}, \\ f_{\mu\rho}^{(m)} &= h_{\mu\rho} + \sum_{\nu\sigma} d_{\nu\sigma}^{(m)}\{2(\mu\rho|\nu\sigma) - (\mu\nu|\rho\sigma)\} \; . \end{aligned} \tag{255}$$

Obviously, the m^{th} Fock matrix can also be computed via the recurrence relation

$$f_{\mu\rho}^{(m)} = f_{\mu\rho}^{(m-1)} + \sum_{\nu\sigma}\{d_{\nu\sigma}^{(m)} - d_{\nu\sigma}^{(m-1)}\}\{2(\mu\rho|\nu\sigma) - (\mu\nu|\rho\sigma)\},$$

i.e. by generating an incremental two-electron repulsion matrix, obtained by contracting the ERIs with an difference density matrix $\Delta\mathbf{d}^{(m)} = \mathbf{d}^{(m)} - \mathbf{d}^{(m-1)}$. Towards convergence, $\Delta\mathbf{d}^{(m)}$ will become very sparse, and thus the prescreening be more and more effective. The advantages of this recursive construction of the Fock matrix can be further enhanced by the 'minimized density difference' approach [137], where rather than simple density differences a *linear combination* of a history of densities (and Fock matrices) is used, which minimizes the density residual. On should note at this point, however, that the prescreening thresholds may have to be tightened towards convergence in order to avoid numerical noise and thus a deterioration of the convergence behaviour of the SCF. Changing the thresholds on the other hand implies the calculation of a full Fock matrix, i.e., a *restart* of the

density difference procedure. Moreover, the DIIS (direct inversion of the iterative subspace [99]) convergence accelerator has to be restarted as well.

The philosophy of the direct SCF approach was based on the observation that the efficiency of integral processing had outgrown the storage and I/O capacities on modern computer systems. Evidently though, after eliminating the storage and I/O bottleneck at the cost of additional CPU time, the evaluation of the ERIs again becomes the bottleneck in large direct SCF calculations, despite of all the ERI prescreening discussed above. Much work has therefore been dedicated to improve the efficiency of ERI evaluation and Fock matrix construction. Some of these ideas can be summarized as *early contraction* schemes, where the Fock matrix is built directly from the two-centre integrals in the Gaussian Product basis (cf. GPT, eq. (248)), avoiding the handling of explicit four-centre ERIs over primitive or contracted basis functions as much as possible. Other ideas go into the direction of (approximately) reexpanding a product of basis functions in a new auxiliary basis (approximate three-centre expansions[138]). The approximate three-centre expansions appear in a different context (RI-DFT, RI-MP2) in other lectures of this winter school. A discussion of these methods is beyond the scope of this brief overview. Excellent overviews of these methods can be found in Refs. [139,140].

6.3 Integral-direct MP2

As shown in section 2.3, the MP2 contribution to the correlation energy for a closed shell system can be written in spin-free formalism as

$$E^{(2)} = \sum_{i,j,a,b} \frac{(ia|jb)[2(ia|jb) - (ib|ja)]}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}, \tag{256}$$

where ϵ_i, ϵ_j, ϵ_a, ϵ_b are the corresponding eigenvalues of the Fock matrix. The MO exchange integrals $(bj|ia)$ are computed from the AO integrals (ERIs) through a four-index transformation as shown in eq. (91). In the following, we will denote the four quarter transformation steps by Q1, Q2, Q3 and Q4, respectively. The nominal operation count (without any prescreening) of the Q1 step scales with $\mathcal{O}(m_{\mathrm{occ}}m^4)$, while the others scale with $\mathcal{O}(m_{\mathrm{occ}}^2 m^3)$, i.e. the cost of all steps increases with $\mathcal{O}(\mathcal{N}^5)$. For applications on large molecules it is therefore essential to reduce this steep scaling by prescreening techniques, similar to the direct SCF case.

The memory requirements of the four individual transformation steps can be minimized by performing these over *fixed shells*. This seems to be quite natural, since the ERIs are generated anyway as individual batches over shell quadruplets. In a straightforward scheme of that type the storage requirements to hold an individual AO ERI batch then are $\mathcal{O}(s^4)$ (s denotes an average shell size, which is independent of the molecular size), $\mathcal{O}(m_{\mathrm{occ}}ms^2)$ for the ERIs after the Q1 and Q2 steps, and $\mathcal{O}(m_{\mathrm{occ}}^2 m^2)$ after the Q3 and Q4 steps, respectively. Apparently, while the computational burden is largest for the initial transformation step, the memory requirements are highest for the final step. In the canonical MP2 case the MO integrals are immediately consumed and accumulated to the MP2 correlation energy, according to eq. (256). A straightforward way to reduce the memory requirements of the critical Q3 and Q4 steps then is to *segment* the first MO index i into indi-

vidual chunks (as large chunks as possible, given by the available memory) and to *multipass* over the AO integral list for each chunk individually [126,127]. This reduces the memory requirements from $\mathcal{O}(m_{\mathrm{occ}}^2 m^2)$ to $\mathcal{O}(I m_{\mathrm{occ}} m^2)$ (I denotes the chunk size) at the cost of repeated ERI evaluations. In order for this algorithm to work, one of the ERI permutational symmetries (i.e. the $(\mu\rho) \leftrightarrow (\nu\sigma)$ symmetry) must be abandoned, thus one integral pass involves twice as many ERIs as the minimal list. The algorithm is free of any I/O operations and can be considered as *fully direct*. Yet the disadvantages are obvious: repeated ERI evaluation might become quite costly, and the number of passes increases quartically with increasing system size and constant memory. A more efficient, *semi direct* algorithm generates in a first step the whole set of *half transformed* integrals $(\mu j|i\nu)$. The transformation of the remaining two indices μ, ν to the virtual basis takes place after an intermediate *bucket sort*, which rearranges the ERIs to *integral matrices* $K_{\mu\nu}^{ij}$, and transforms individual $\mathbf{K}^{ij}$ matrices one after the other. If the permutational symmetry of the slow pair $(\mu\rho)$ (i.e. $\mu \leftrightarrow \rho$) is abandoned, the maximum memory requirements are solely $\mathcal{O}(s m_{\mathrm{occ}} m^2)$. Such an algorithm is outlined in pseudocode below (algorithm A)

```
DO M=1,NShell
  DO R=1,NShell
    DO N=1,M
      DO S=1,N | R (for M=N)
        Compute integral block (MR|NS)
        Q1 step over shell block:
        (MR|Nj) = (MR|Nj) + (MR|NS) * X(S,j)
        (MR|Sj) = (MR|Sj) + (MR|NS) * X(N,j)
      END DO
    END DO
    (Mi|Nj) = (Mi|Nj) + (MR|Nj) * X(R,i)
  END DO
  write (Mi|Nj) to disk
END DO
perform bucket sort/(Mi|Nj)=(Mi|Nj)+(Nj|Mi)
```

Note, that in order to keep the $(\mu\rho) \leftrightarrow (\nu\sigma)$ permutational symmetry the triangularity in the operator indices i, j is lost. The final operator matrices $\mathbf{K}^{ij}(i \geq j)$ are formed by adding up the partial results $\mathcal{K}_{\mu\nu}^{ij} + \mathcal{K}_{\nu\mu}^{ji}(i \geq j)$, which is performed during the bucket sort, as indicated above.
By virtue of an elaborate paging algorithm, it is even possible to maintain also the $\mu \leftrightarrow \rho$ permutational symmetry (algorithm B), i.e.

```
    R_End=0
    R_Pass=0
1   R_Start=R_End+1
    R_End=MIN(NShell,R_End+R_Batch)
    R_Pass=R_Pass+1
```

```
if(R_Pass.gt.1) Read (Ri|Nj) for shells R_Start to R_End
DO M=R_Start,NShell
  IF(R_Pass.gt.1.and.M.gt.R_End) Read (Mi|Nj) for shell M
  DO R=R_Start,MIN(R_End,M)
    DO N=1,M
      DO S=1,N | R (for M=N)
        Compute integral block (MR|NS)
        Q1 step over shell block:
        (MR|Nj) = (MR|Nj) + (MR|NS) * X(S,j)
        (MR|Sj) = (MR|Sj) + (MR|NS) * X(N,j)
      END DO
    END DO
    (Mi|Nj) = (Mi|Nj) + (MR|Nj) * X(R,i)
    (Ri|Nj) = (Ri|Nj) + (MR|Nj) * X(M,i)
  END DO
  IF(M.GT.R_End) Write (Mi|Nj) for shell M
END DO
Write (Ri|Nj) for shells R_Start to R_End
If(R_End.LT.NShell) goto 1
perform bucket sort/(Mi|Nj)=(Mi|Nj)+(Nj|Mi)
```

This means that the full permutational symmetry of the AO ERIs is exploited. This algorithm is very efficient for molecular systems of intermediate size. However, for large systems and limited memory, the paging overhead might become too excessive (even though no multipassing whatsoever over the integral list is involved, as in the fully direct scheme), and algorithm A becomes more efficient.

The Q1 and Q2 transformation steps require matrix multiplications, in which at least one of the matrix dimensions corresponds to the shell size. For small shells the vector lengths are too short for a good performance to be achieved. Therefore, it is advantageous to merge adjacent `R` and `S` shells until an upper limit of 32-64 basis functions is reached. Significant speedups (factors of 4-6) were observed, if such *shell merging* was invoked [60].

For applications on larger molecules, integral prescreening is of utmost importance. In order to assess the values of the AO ERIs, the Schwartz inequality (eq. 251) is again employed. Furthermore a test density $\mathbf{D}^{\max}$ is constructed from the MO coefficient matrix $\mathbf{C}$ as

$$D_{\rho\sigma}^{\max} = \underset{ij}{\mathrm{Max}} C_{\sigma i} C_{\rho j} \tag{257}$$

The prescreening criterions for the direct transformation at the level of shell quadrulets then are

$$Q_{MR} Q_{NS} D_{RS}^{\max} \geq \tau_1 \tag{258}$$

before integral evaluation, and

$$\underset{\mu\in M,\rho\in R,\nu\in N,\sigma\in S}{\mathrm{Max}} (\mu\rho|\nu\sigma) D_{RS}^{\max} \geq \tau_2 \tag{259}$$

before the Q1 step, respectively. Such a prescreening leads to a reduction of the computational cost of the dominant Q1 step from $\mathcal{O}(\mathcal{N}^5)$ to $\mathcal{O}(\mathcal{N}^3)$ [60]. The overall scaling however deteriorates again for larger molecules due to the subsequent transformations steps, which, because of the delocalized character of canonical orbitals, scale worse than $\mathcal{O}(\mathcal{N}^3)$. In particular the Q4 step (i.e. the transformation of the $\mathbf{K}^{ij}_{\mu\nu}$ to the canonical virtuals) would still scale as $\mathcal{O}(\mathcal{N}^5)$, although with a small prefactor, but nevertheless will ultimately constitute the bottleneck of the calculation. The remedy to this problem are *local correlation methods*, discussed in section 4. In combination with local correlation methods integral-direct MP2 algorithms with linear cost scaling have been implemented, which enable calculations of molecules with more than 2000 basis functions and 500 correlated electrons [54].

6.4 Integral-direct MCSCF

In MCSCF calculations the orbitals are optimized simultaneously with the CI coefficients. Thus, an integral transformation is required in each iteration, which constitutes one of the major bottlenecks in conventional MCSCF calculations. In a direct scheme, this bottleneck is even much more severe, since each direct transformation also involves recomputation of all AO ERIs. It is therefore of utmost importance that the MCSCF converges in as few iterations as possible.

MCSCF orbital optimization methods can be classified as first-order or second-order methods. In the former only the first derivatives of the energy with respect to the variational parameters are computed exactly, and updates of the parameters are obtained using some approximation of the Hessian (e.g. a BFGS update scheme). In first-order methods the coupling of the orbitals and CI-coefficients is neglected. One particular advantage of first-order methods is that only a very compact set of transformed integrals is required, i.e. an integral distribution of the form $(pj|kl)$ with only a single external index. In fact, j, k, l here run just over active orbitals, while the inactive orbitals (doubly occupied in all CSFs) can be accounted for by a single Fock matrix [141,142]. Thus, any storage bottleneck connected to the integral transformation is avoided. An integral-direct first-order MCSCF method has been described by Frisch *et al.*[124].

In second-order methods, also the second energy derivatives are computed exactly, yielding quadratic convergence near the final solution. Naturally, first-order methods require less effort per iteration, but are often slowly convergent and appear to be only useful for the optimization of CASSCF wavefunctions[141]. In this case convergence is facilitated by the fact that orbital rotations among active orbitals are redundant. Even with second-order methods convergence is often difficult to achieve for general MCSCF wavefunctions[142]. The radius of convergence and the speed of convergence can be substantially increased by taking into account certain higher-order terms, as first proposed by Werner and Meyer[143,144] and further refined by Werner and Knowles[98,91]. Using the latter method (in the following denoted WMK), convergence can often be achieved in only 2-3 iterations, in particular for CASSCF wavefunctions. Almost cubic convergence behaviour is observed near the solution. In the light of the discussion above, the WMK method is particulary useful in an integral-direct context, while the advantage of the simple and efficient

transformation of first-order methods is spoilt by its slow convergence behaviour. The integral sets required by the WMK method are identical to those used by ordinary second-order methods: in additon to the exchange integrals $(ip|jq)$ also the Coulomb integrals $(pq|ij)$ are necessary. Furthermore, the very same integral sets, generated in the last iteration, can be reused in a subsequent CASPT2 or MRCI calculation. The additional Coulomb integral set can be produced simultaneously with the exchange integrals by modifying the above MP2 transformation algorithm A in the following way (algorithm C):

```
DO M=1,NShell
  DO R=1,NShell
    DO N=1,NShell
      DO S=1,R
        Compute integral block (MR|NS)
        Q1 step over shell block:
        (MR|Nj) = (MR|Nj) + (MR|NS) * X1(S,j)
        (MR|Sj) = (MR|Sj) + (MR|NS) * X1(N,j)
      END DO
    END DO
    Q2 (J) step:
    (MR|ij) = (MR|Nj) * X2(N,i)    (summed over N)
    write (MR|ij) to disk
    Q2 (K) step:
    (Mi|Nj) = (Mi|Nj) + (MR|Nj) * X2(R,i)
  END DO
  write (Mi|Nj) to disk
END DO
perform bucket sort
```

Note, that compared to algorithm A the permutational symmetry between the pairs $(\mu\rho) \leftrightarrow (\nu\sigma)$ is lost, thus the AO integral list in algorithm C is four times as long as the minimal list. As in the MP2 case (algorithm B), the permutational symmetry $(\mu\rho)$ can be maintained by using an analogous paging algorithm, which might be advantageous for intermediate cases.

6.5 Integral-direct multireference correlation methods

The internally contracted MRCI and MRPT methods as discussed in section 5 can be formulated in terms of matrix operations[142] involving the same Coulomb and exchange matrices $\mathbf{J}^{ij}$ and $\mathbf{K}^{ij}$ as needed in the preceeding MCSCF. In the MRCI and MRPT3 all contributions of 4-external integrals $(ab|cd)$ can be taken into account by computing for each pair P an *external exchange operator* (EEO), as defined in eq. (106)[117,118,108]. These operators can be computed directly from the two-electron integrals in the AO basis by first transforming the amplitude matrices into the AO basis and finally transforming these back into the MO basis (cf. eqs (107). For an integral-direct implementation the internally contracted MRCI

scheme is particularly useful, since the number of pairs and thus external exchange operators that need to be computed is minimized and does not depend on the number of reference configurations. In uncontracted MRCI methods the number of pairs P for which the EEOs $\mathbf{K}(\mathbf{T}^P)$ must be computed is excessively larger than in the internally contracted case. This does not only lead to higher computational cost, but also to a storage bottleneck in the direct evaluation of these operators

The direct construction of the EEOs from the minimal AO integral list is accomplished by contracting two indices of the AO ERI $(\mu\rho|\sigma\nu)$ with the two AO indices of the backtransformed amplitudes 107, in all possible ways, which result in exchange type contributions, and can be regarded as a 'Fock build' (excluding Coulomb contributions) of n_P Fock matrices simultaneously (n_P denotes the number of pairs P). A shell driven out-of-core algorithm for such a construction of the EEOs, as implemented in MOLPRO [60], is given in pseudocode below (module `DKEXT`).

```
    R_End=0
    R_Pass=0
1   R_Start=R_End+1
    R_End=MIN(NShell,R_End+R_Batch)
    Read amplitudes for shells R_Start to R_End
    R_Pass=R_Pass+1
    IF(R_Pass.gt.1) Read operators for shells R_Start to R_End
    DO M=R_Start,NShell
      If(M.GT.R_End) then
        Read amplitudes for shell M
        If(R_Pass.gt.1)  Read operators for shell M
      End If
      DO R=R_Start,MIN(R_End,M)
        DO N=1,M
          S_End=N
          If(N.EQ.M) S_End=R
          DO S=1,S_End
            Compute integral block (MR|NS)
            Compute contributions to operators
          END DO
        END DO
      END DO
      IF(M.GT.R_End) Write operators for shell M
    END DO
    Write operators for shells R_Start to R_End
    If(R_End.LT.NShell) goto 1
```

The algorithm employs a paging algorithm, which is quite similar to that used in the direct transformation scheme discussed in section 6.3. The amplitudes and EEOs are presorted according to shell blocks $T^{MR}_{P,\mu\rho}$ with M running slowest, and stored on disk. In this way it is possible to read/write them for a given shell M

and for all P and R.

All contributions arising from integrals over one occupied and three external orbitals $(ia|bc)$ can be taken into account by an additional set of EEOs $\mathbf{K}(\mathbf{D}^P)$, where $\mathbf{D}^P$ are modified coefficient matrices[117,118,22], which differ from the $\mathbf{T}^P$ by the addition of internal-external blocks arising from contributions of single excitations. In single-reference methods (CISD, MP4(SDQ), QCSID, CCSD) as well as for evaluating the MRPT3 energy it is sufficient to compute only the latter set of operators. In MRCI calculations, this would in principle be possible as well, but since then complicated correction terms are necessary[118] it is easier to compute the operators $\mathbf{K}(\mathbf{T}^P)$ and $\mathbf{K}(\mathbf{D}^P)$ separately. Of course, the two sets can be computed together in a single integral pass.

Since the EEOs depend explicitly on the amplitudes that must be computed in each iteration. The computational complexity of EEO formation is nominally a task $\mathcal{O}(m_{\rm occ}^2 m^4) = \mathcal{O}(\mathcal{N}^6)$. In an integral-direct context this can be reduced to $\approx \mathcal{O}(\mathcal{N}^4)$ by virtue of integral prescreening [60]. In order to get efficient prescreening, it is important to include the amplitudes into the prescreening scheme. Nevertheless, in integral-direct calculations with large basis sets, the EEO construction often dominates the computational effort.

6.6 Integral-direct coupled cluster methods

The first integral-direct CCSD method was developed by Koch and coworkers[129,145]. In this method the transformed integrals are never stored on disk. Instead, "distributions" of AO integrals $(\mu\rho|\nu\sigma)$ are generated for fixed μ, all ρ, $\nu \geq \sigma$. One such distribution at a time is kept in memory and consumed immediately to compute all contributions to the CCSD residual (fully direct CCSD). This method, although very efficient on vector computers due to long vector lenghths, suffers from some severe bottlenecks (most importantly, the m^3 memory requirements of the integral distributions, mentioned above), which limit the application range for larger systems. An alternative method has been proposed by Schütz, Werner and Lindh [60], which differs from the above method by the fact, that the partially transformed integrals are stored on disk ($3/2 m_{\rm occ}^2 m^2$ words are required). Considering that the doubles amplitudes as the variational parameters of the iterative CCSD procedure and the residuals have to be stored on disk anyway in several instances (due to DIIS convergence acceleration), with a required diskspace of $n_{\rm DIIS} m_{\rm occ}^2 m^2$, this certainly does not constitute a further bottleneck, and seems to be a reasonable strategy. The immediate advantage is that the remaining program remains entirely unchanged, and that the same integral-direct modules as for the MCSCF and MRCI programs can be used. Furthermore, in such a scheme the maximum memory requirements can be reduced to $\mathcal{O}(m_{\rm occ} m^2)$, and to $\mathcal{O}(\mathcal{N})$ for local CCSD (cf. section 4.2).

The MP3, MP4(SDQ), QCISD and CCSD methods, which all are related, require the same internal operators $\mathbf{J}^{ij}$, $\mathbf{K}^{ij}$, and the EEOs $\mathbf{K}(\mathbf{D}^{ij})$ as introduced for the MRCI case in the previous section. A further complication arises in the CCSD method [22], where the additional operators $\mathbf{J}(\mathbf{E}^{ij})$ and $\mathbf{K}(\mathbf{E}^{ij})$ (cf. eqs. (132)) are needed. As discussed in section 2.5, these operators can be obtained by a generalized integral transformation (cf. eqs. (133)-(137)). This transformation can be

performed using the same integral-direct module as employed for generating the $\mathbf{J}^{ij}$ and $\mathbf{K}^{ij}$ matrices, but since they depend on the singles amplitudes they must be performed in each iteration. An important point to notice is that the latter operators are only needed for CCSD, but not for the QCISD (quadratic configuration interaction) method[33]. While the computational effort for these two methods is not too much different in conventional calculations[22], in the integral-direct case the full CCSD takes significantly more time, due to this additional transformation which must be performed in each iteration. For most applications, QCISD and CCSD results are very similar, and QCISD may often be more cost effective for integral-direct calculations of large molecules, even though from a theoretical point of view CCSD is more satisfactory. If the 3-external integrals are available though, as is usually the case for local CCSD calculations, then the construction of the $\mathbf{J}(\mathbf{E}^{ij})$ and $\mathbf{K}(\mathbf{E}^{ij})$ operators takes little time, hence there is little reason to use the QCISD model in that case.

Acknowledgments

Financial support from the EC as part of the TMR network "Potential Energy Surfaces for Spectroscopy and Dynamics", contract No. FMRX-CT96-088 (DG 12 – BIUO) and as part of the RTN network "Theoretical Studies of Electronic and Dynamical Processes in Molecules and Clusters (THEONET II)", contract No. RTN1-1999-00121 is gratefully acknowledged. Much of the research described in these notes has been supported by DFG, EPSRC, Fonds der Chemischen Industrie and BASF AG.

References

1. R. H. Nobes, J. A. Pople, L. Radom, N. C. Handy, and P. J. Knowles, Chem. Phys. Letters **138**, 481 (1987).
2. P. J. Knowles and N. C. Handy, J. Phys. Chem. **92**, 3097 (1988).
3. E. R. Davidson, J. Comput. Phys. **17**, 87 (1975).
4. T. Kato, Commun. Pure Appl. Math. **10**, 151 (1957).
5. R. T. Pack and W. Byers Brown, J. Chem. Phys. **45**, 556 (1966).
6. W. A. Bingel, Z. Naturforsch. Teil A **18**, 1249 (1963).
7. V. A. Rassolov and D. M. Chipman, J. Chem. Phys. **104**, 9908 (1996).
8. E. A. Hylleraas, Z. Phys. A **54**, 347 (1929).
9. H. M. James and A. S. Coolidge, J. Chem. Phys. **1**, 825 (1933).
10. W. Kutzelnigg, Theor. Chim. Acta **68**, 445 (1985).
11. W. Kutzelnigg and W. Klopper, J. Chem. Phys. **94**, 1985 (1991).
12. D. C. Clary and N. C. Handy, Phys. Rev. **A14**, 1607 (1976).
13. D. C. Clary, Mol. Phys. **34**, 793 (1977).
14. P. Jørgensen and J. Simons, *Second Quantization–Based Methods in Quantum Chemistry* (Academic Press, New York, 1981).
15. J. Almlöf and P. R. Taylor, J. Chem. Phys. **86**, 4070 (1987).
16. J. Almlöf and P. R. Taylor, J. Chem. Phys. **92**, 551 (1990).
17. P.-O. Widmark, P.-Å. Malmqvist, and B. O. Roos, Theor. Chim. Acta **77**,

291 (1990).
18. J. Almlöf and P. R. Taylor, Adv. Quant. Chem. **22**, 301 (1991).
19. T. H. Dunning Jr., J. Chem. Phys. **90**, 1007 (1989).
20. W. Meyer, J. Chem. Phys. **64**, 2901 (1976).
21. P. Pulay, S. Saebø, and W. Meyer, J. Chem. Phys. **81**, 1901 (1984).
22. C. Hampel, K. A. Peterson, and H.-J. Werner, Chem. Phys. Lett. **190**, 1 (1992).
23. W. Meyer, Int. J. Quantum Chem. Symp. **5**, 341 (1971).
24. W. Meyer, J. Chem. Phys. **58**, 1017 (1973).
25. R. Ahlrichs, P. Scharf, and C. Ehrhardt, J. Chem. Phys. **82**, 890 (1985).
26. S. R. Langhoff and E. R. Davidson, Int. J. Quant. Chem. **8**, 61 (1974).
27. B. O. Roos and P. E. M. Siegbahn, in *Methods of Electronic Structure Theory*, edited by H. F. Schaefer III (Plenum, New York, 1977).
28. J. Čížek, J. Chem. Phys. **45**, 4256 (1966).
29. J. Čížek, Adv. Chem. Phys. **14**, 35 (1969).
30. J. Čížek and J. Paldus, Int. J. Quantum Chem. **5**, 359 (1971).
31. G. D. Purvis and R. J. Bartlett, J. Chem. Phys. **76**, 1910 (1982).
32. G. E. Scuseria, C. L. Janssen, and H. F. Schaefer III, J. Chem. Phys. **89**, 7382 (1988).
33. J. A. Pople, M. Head-Gordon, and K. Raghavachari, J. Chem. Phys. **87**, 5968 (1987).
34. J. Čížek and J. Paldus, Phys. Scripta **21**, 251 (1980).
35. R. J. Bartlett and G. D. Purvis, Phys. Scripta **21**, 255 (1980).
36. R. A. Chiles and C. E. Dykstra, J. Chem. Phys. **74**, 4544 (1981).
37. G. Scuseria and H. F. Schaefer III, Chem. Phys. Lett. **142**, 354 (1987).
38. N. C. Handy, J. A. Pople, M. Head-Gordon, K. Raghavachari, and G. W. Trucks, Chem. Phys. Lett. **164**, 185 (1989).
39. K. Raghavachari, J. A. Pople, E. S. Replogle, M. Head-Gordon, and N. C. Handy, Chem. Phys. Lett. **167**, 115 (1990).
40. M. Urban, J. Noga, S. J. Cole, and R. J. Bartlett, J. Chem. Phys. **83**, 4041 (1985).
41. K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chem. Phys. Letters **157**, 479 (1989).
42. S. A. Kucharski and R. J. Bartlett, Adv. Quantum Chem. **18**, 281 (1986).
43. S. A. Kucharski, J. Noga, and R. J. Bartlett, J. Chem. Phys. **90**, 7282 (1989).
44. K. Raghavachari, J. A. Pople, E. S. Replogle, and M. Head-Gordon, J. Phys. Chem. **94**, 5579 (1990).
45. Z. He and D. Cremer, Theor. Chim. Acta **85**, 305 (1993).
46. M. J. O. Deegan and P. J. Knowles, Chem. Phys. Letters **227**, 321 (1994).
47. P. J. Knowles, C. Hampel, and H.-J. Werner, J. Chem. Phys. **99**, 5219 (1993).
48. C. Janssen and H. F. Schaefer III, Theor. Chim. Acta **79**, 1 (1991).
49. P. J. Knowles, C. Hampel, and H.-J. Werner, J. Chem. Phys. **111**, 0000 (2000).
50. P. Neogrády, M. Urban, and I. Hubač, J. Chem. Phys. **100**, 3706 (1994).
51. P. G. Szalay and J. Gauss, J. Chem. Phys. **107**, 9028 (1997).
52. S. Saebø and P. Pulay, Annu. Rev. Phys. Chem. **44**, 213 (1993).

53. C. Hampel and H.-J. Werner, J. Chem. Phys. **104**, 6286 (1996).
54. M. Schütz, G. Hetzer, and H.-J. Werner, J. Chem. Phys. **111**, 5691 (1999).
55. R. A. Friesner, R. B. Murphy, M. D. Beachy, M. N. Ringnalda, W. T. Pollard, B. D. Dunietz, and Y. Cao, J. Phys. Chem. A **103**, 1913 (1999).
56. G. Reynolds, T. J. Martinez, and E. A. Carter, J. Chem. Phys. **105**, 6455 (1996).
57. P. Pulay, Chem. Phys. Letters **100**, 151 (1983).
58. P. Pulay and S. Saebø, Theor. Chim. Acta **69**, 357 (1986).
59. S. Saebø and P. Pulay, J. Chem. Phys. **86**, 914 (1987).
60. M. Schütz, R. Lindh, and H.-J. Werner, Mol. Phys. **96**, 719 (1999).
61. M. Schütz and H.-J. Werner, manuscript in preparation.
62. M. Schütz and H.-J. Werner, Chem. Phys. Lett., in press.
63. S. F. Boys, in *Quantum Theory of Atoms, Molecules, and the Solid State*, edited by P. O. Löwdin, page 253 (Academic, New York, 1966).
64. C. Edmiston and K. Ruedenberg, J. Chem. Phys. **43**, S97 (1965).
65. J. Pipek and P. G. Mezey, J. Chem. Phys. **90**, 4916 (1989).
66. M. Head-Gordon, P. E. Maslen, and C. A. White, J. Chem. Phys. **108**, 616 (1998).
67. G. E. Scuseria and P. Y. Ayala, J. Chem. Phys. **111**, 8330 (1999).
68. J. W. Boughton and P. Pulay, J. Comput. Chem. **14**, 736 (1993).
69. G. Hetzer, P. Pulay, and H.-J. Werner, Chem. Phys. Lett. **290**, 143 (1998).
70. G. Rauhut, P. Pulay, and H.-J. Werner, J. Comput. Chem. **19**, 1241 (1998).
71. A. ElAzhary, G. Rauhut, P. Pulay, and H.-J. Werner, J. Chem. Phys. **108**, 5185 (1998).
72. M. Schütz, G. Rauhut, and H.-J. Werner, J. Phys. Chem. A **102**, 5997 (1998).
73. N. Runeberg, M. Schütz, and H.-J. Werner, J. Chem. Phys. **110**, 7210 (1999).
74. J. Gauss and H.-J. Werner, Mol. Phys., in press.
75. MOLPRO is a package of *ab initio* programs written by H.-J. Werner and P. J. Knowles, with contributions from J. Almlöf, R. D. Amos, A. Berning, P. Celani, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, S. T. Elbert, C. Hampel, G. Hetzer, T. Korona, R. Lindh, A. W. Lloyd, W. Meyer, M. E. Mura, A. Nicklass, K. Peterson, R. Pitzer, P. Pulay, G. Rauhut, M. Schütz, H. Stoll, A. J. Stone, P. R. Taylor, and T. Thorsteinsson.
76. J. Paldus, J. Chem. Phys. **61**, 5321 (1974).
77. J. Hinze, editor, *The Unitary Group* (Springer–Verlag, Berlin, 1979).
78. E. R. Davidson, in *Methods in Computational Molecular Physics*, edited by G. H. F. Diercksen and S. Wilson (Reidel, Dordrecht, 1983).
79. P. E. M. Siegbahn, Chem. Phys. Letters **109**, 417 (1984).
80. J. Olsen, B. O. Roos, P. Jørgensen, and H. J. A. Jensen, J. Chem. Phys. **89**, 2185 (1988).
81. P. J. Knowles and N. C. Handy, Chem. Phys. Letters **111**, 315 (1984).
82. P. J. Knowles and N. C. Handy, Comput. Phys. Commun. **54**, 75 (1989).
83. S. Zarrabian, C. R. Sarma, and J. Paldus, Chem. Phys. Letters **155**, 183 (1989).
84. D. M. Brink and G. R. Satchler, *Angular Momentum* (Clarendon, Oxford, 2nd edition, 1968).

85. R. N. Zare, *Angular Momentum* (Wiley, New York, 1988).
86. R. Pauncz, *Spin Eigenfunctions* (Plenum, New York, 1979).
87. P. J. Knowles and H.-J. Werner, Chem. Phys. Lett. **145**, 514 (1988).
88. K. Ruedenberg, L. M. Cheung, and S. T. Elbert, Int. J. Quantum Chem. **16**, 1069 (1979).
89. B. O. Roos, P. Taylor, and P. E. M. Siegbahn, Chem. Phys. **48**, 157 (1980).
90. P. E. M. Siegbahn, J. Almlöf, A. Heiberg, and B. O. Roos, J. Chem. Phys. **74**, 2384 (1981).
91. P. J. Knowles and H.-J. Werner, Chem. Phys. Letters **115**, 259 (1985).
92. M. R. Hoffmann, D. J. Fox, J. F. Gaw, Y. Osamura, Y. Yamaguchi, R. S. Grev, G. Fitzgerald, H. F. Schaefer III, P. J. Knowles, and N. C. Handy, J. Chem. Phys. **80**, 2660 (1984).
93. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran 77: The Art of Scientific Computing* (Cambridge University Press, 2nd edition, 1992).
94. P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization* (Academic Press, 1981).
95. J. A. Pople, R. Krishnan, H. B. Schlegel, and J. S. Binkley, Int. J. Quant. Chem. **S13**, 225 (1979).
96. J. Olsen, D. L. Yeager, and P. Jørgensen, Adv. Chem. Phys. **54**, 1 (1983).
97. D. Yarkony, Chem. Phys. Letters **77**, 634 (1981).
98. H.-J. Werner and P. J. Knowles, J. Chem. Phys. **82**, 5053 (1985).
99. P. Császár and P. Pulay, J. Mol. Struc. **114**, 31 (1984).
100. C. Møller and M. S. Plesset, Phys. Rev. **46**, 618 (1934).
101. J. A. Pople, R. Krishnan, H. B. Schlegel, and J. S. Binkley, Int. J. Quant. Chem. **14**, 545 (1978).
102. R. J. Bartlett and D. M. Silver, J. Chem. Phys. **62**, 3258 (1975).
103. R. B. Murphy and R. P. Messmer, Chem. Phys. Letters **183**, 443 (1991).
104. K. Andersson, P.-Å. Malmqvist, and B. O. Roos, J. Chem. Phys. **96**, 1218 (1992).
105. K. Hirao, Chem. Phys. Letters **196**, 397 (1992).
106. B. O. Roos, K. Andersson, M. P. Fulscher, P. A. Malmqvist, L. Serranoandres, K. Pierloot, and M. Merchan, Adv. Chem. Phys. **93**, 219 (1996).
107. P. Celani and H.-J. Werner, J. Chem. Phys., in press.
108. H.-J. Werner, Mol. Phys. **89**, 645 (1996).
109. R. J. Buenker and S. D. Peyerimhoff, Theor. Chim. Acta **35**, 33 (1974).
110. P. E. M. Siegbahn, Int. J. Quantum Chem. **18**, 1229 (1980).
111. J. Lischka, R. Shepard, F. B. Brown, and I. Shavitt, Int. J. Quantum Chem. Symp. **15**, 91 (1981).
112. V. R. Saunders and J. H. van Lenthe, Mol. Phys. **48**, 923 (1983).
113. W. Meyer, in *Methods of Electronic Structure Theory*, edited by H. F. Schaefer III (Plenum, New York, 1977).
114. P. E. M. Siegbahn, J. Chem. Phys. **72**, 1647 (1980).
115. C. W. Bauschlicher, P. R. Taylor, N. C. Handy, and P. J. Knowles, J. Chem. Phys. **85**, 1469 (1986).
116. C. W. Bauschlicher, Jr., S. R. Langhoff, and P. R. Taylor, Adv. Chem. Phys.

77, 103 (1990).
117. H.-J. Werner and E. A. Reinsch, J. Chem. Phys. **76**, 3144 (1982).
118. H.-J. Werner and P. J. Knowles, J. Chem. Phys. **89**, 5803 (1988).
119. E. R. Davidson, in *The world of quantum chemistry*, edited by R. Daudel and B. Pullman (Reidel, Dordrecht, 1974).
120. R. J. Gdanitz and R. Ahlrichs, Chem. Phys. Lett. **143**, 413 (1988).
121. P. G. Szalay and R. J. Bartlett, Chem. Phys. Letters **214**, 481 (1993).
122. J. Almlöf, J. K. Faegri, and K. Korsell, J. Comput. Chem. **3**, 385 (1982).
123. P. Taylor, Int. J. Quantum Chem. **31**, 521 (1987).
124. M. Frisch, I. N. Raganzos, M. A. Robb, and H. B. Schlegel, Chem. Phys. Lett. **189**, 524 (1992).
125. S. Sæbø and J. Almlöf, Chem. Phys. Letters **154**, 83 (1989).
126. M. Head-Gordon, J. Pople, and M. Frisch, Chem. Phys. Letters **153**, 503 (1988).
127. M. Schütz and R. Lindh, Theor. Chim. Acta **95**, 13 (1997).
128. M. Frisch, M. Head-Gordon, and J. Pople, Chem. Phys. Letters **166**, 275 (1990).
129. H. Koch, O. Christiansen, R. Kobayashi, P. Jørgensen, and T. Helgaker, Chem. Phys. Lett. **228**, 233 (1994).
130. W. Klopper and J. Noga, J. Chem. Phys. **103**, 6127 (1995).
131. R. Lindh, in *The Encyclopedia of Computational Chemistry Vol.2*, edited by P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. S. III, and P. R. Schreiner, page 1337 (John Wiley & Sons: Chichester, 1998).
132. C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Letters **230**, 8 (1994).
133. C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Letters **253**, 268 (1996).
134. J. C. Burant, G. E. Scuseria, and M. J. Frisch, J. Chem. Phys. **105**, 8969 (1996).
135. M. Challacombe, E. Schwegler, and J. Almlöf, J. Chem. Phys. **104**, 4685 (1996).
136. C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
137. M. Häser and R. Ahlrichs, J. Comput. Chem. **10**, 104 (1989).
138. O. Vahtras, J. Almlöf, and M. Feyereisen, Chem. Phys. Letters **213**, 514 (1993).
139. J. Almlöf, in *Modern Electronic Structure Theory*, edited by D. Yarkony, number I in Advanced Series in Physical Chemistry - Vol. 2, page 110 (World Scientific Publishing Co. Pte. Ltd., 1995).
140. J. Almlöf, in *Lecture Notes in Quantum Chemistry, European Summer School in Quantum Chemistry*, edited by B. Roos, number 64 in Lecture Notes in Chemistry, page 1 (Springer-Verlag Berlin Heidelberg, 1994).
141. B. O. Roos, Int. J. Quantum Chem. Symp. **14**, 175 (1980).
142. H.-J. Werner, Adv. Chem. Phys. **69**, 1 (1987).
143. H.-J. Werner and W. Meyer, J. Chem. Phys. **73**, 2342 (1980).
144. H.-J. Werner and W. Meyer, J. Chem. Phys. **74**, 5794 (1981).

145. H. Koch, A. S. de Merás, T. Helgaker, and O. Christiansen, J. Chem. Phys. **104**, 4157 (1996).